A Morphological Lexicon of Esperanto with Morpheme Frequencies

Eckhard Bick
University of Southern Denmark
Campusvej 55, DK-5230 Odense M
Email: eckhard.bick@mail.dk

Abstract

This paper discusses the internal structure of complex Esperanto words (CWs). Using a morphological analyzer, possible affixation and compounding is checked for over 50,000 Esperanto lexemes against a list of 17,000 root words. Morpheme boundaries in the resulting analyses were then checked manually, creating a CW dictionary of 28,000 words, representing 56.4% of the lexicon, or 19.4% of corpus tokens. The error percentage of the EspGram morphological analyzer for new corpus CWs was 4.3% for types and 6.4% for tokens, with a recall of almost 100%, and wrong/spurious boundaries being more common than missing ones. For pedagogical purposes a morpheme frequency dictionary was constructed for a 16 million word corpus, confirming the importance of agglutinative derivational morphemes in the Esperanto lexicon. Finally, as a means to reduce the morphological ambiguity of CWs, we provide POS likelihoods for Esperanto suffixes.

Keywords: Morphological Analysis, Esperanto, Affixation, Compounding, Morpheme Frequencies

1. Introduction

As an artificial language with a focus on regularity and facilitation of language acquisition, Esperanto was designed with a morphology that allows (almost) free, productive combination of roots, affixes and inflexion endings. Thus, the root *'san'* (healthy) not only accepts its prototypical adjectival ending *'-a'*, but also other part-of-speech endings *('san|e'* - healthily, *'san|on'* - cheers)[1], as well as prefixes *('mal+san|a'* - unhealthy [mal=opposite]), suffixes *('sanigi'* - cure [ig=make]) or multiple affixes *('mal+san+ul+ej|o'* - hospital [ul=person, ej=place]). In addition, compounding is common *('san+serv|oj'* - health services, *'san+asekur|o'* - health insurance). This morphological versatility reduces the number of to-be-learned lexemes and is - together with high linguistic transparency - generally cited as a main reason for the language's easiness as an L2 and its usefulness as a propedeutic language (e.g. Telier 2013). Given the agglutinative properties of the language, it is pedagogically interesting to know which morphemes contribute most to the Esperanto lexicon, in order to teach them first. In other words, on top of basic *word* lists and *word* frequencies (Quasthoff et al. 2014), Esperanto teachers and text book authors also need *morpheme* frequencies. The work described here is intended to provide just that - a frequency dictionary of Esperanto morphemes, as well as the tools required to build it from corpus data.

2. Morphological analysis

From a language technology perspective, inflexional regularity, morphological transparency and surface-based access to semantic features turn POS tagging of Esperanto into a non-task, and facilitate the parsing of syntactic and semantic structures (Bick 2007). As a low-level, "local" NLP task, one would expect morphology, in particular, to be a simple task for the same reasons. However, what is transparent for a human beholder, is not necessarily as easy to grasp for a computer program. Thus, while POS is completely unambiguous in Esperanto, unrestricted compounding and affixation do produce a fair amount of theoretical morphological ambiguity. Early computational work in the area includes two-level morphology (Hana 1998), but suffered from lexicographical limitations and a lack of disambiguation, and application-driven analyzers (spell checking, machine translation) typically regard the task only as an add-on for out-of-lexicon words (e.g. Hun-Spell affix files, Blahuš 2009). This can be true also of parsers, due to their focus on syntax. Thus, the analyzer stage of the EspGram parser (Bick 2009) used in our experiments has a large lexicon and does handle some disambiguation, but it originally only performed morphological analysis of unknown words, or where necessary to predict syntactic-semantic features that it could not look up in its dictionary. In this paper we will describe lexicographical work needed to assign internal structure to *all* complex words (CWs), both lexicalized and productive, and resolve any arising ambiguity, in order to achieve reliable and complete decomposition of Esperanto words for our morpheme frequency dictionary.

Three types of such ambiguity can be distinguished: The first concerns simplex words that also have an analytical reading (1), the second is ambiguity between two possible

[1] The POS markers, attached agglutinatively at the end, after possible suffixes, are the following: -o = noun, -a = adjective, -e = adverb, -i = verb (infinitive). After these, a plural '-j' (for -a/o) and accusative '-n' (for -a/o/e) can be added as an inflection ending. For verbs, the infinitve ending can be replaced by a tense vowel (a = present, i = past, o = future) plus a finity ending '-s' (e.g. *aĉetis* [bought]). Participles are built by adding '-t' (passive) or '-nt' (active) instead (e.g. *aĉetinta* [having bought]).

compound (+)[2] or affix (%) cuts (2), and in the third, at least 1 part of a complex word is itself a compound/affixation (3).

1a) '*insekto*' (insect) - *'in+sekto'* (feminist [female] sect)
1b) 'genetiko' (genetics) - *'gen+etiko'* (gene ethics)
2a) 'bov+okulo' (cow eye) -*'bov~o-kulo'* (cow mosquito)
2b) 'el+flui' (flow out) - *'elf-lui' (*rent elves)
2c) 'martir%igo' (martyrization) - *'mar+tir%igo'* (sea pulling) - *'mart+ir%igo'* (March walking)
3a) 'ĉef+staci|domo' (main railway station) - *'ĉef|staci+domo'* (headquarter house)
3b) *'mal+ver+ŝajna* (un-likely) vs. *'malv+er+ŝajna'* (mallow-part resembling)

Besides a good algorithm for multiple derivation, an automatic morphological analyzer for Esperanto needs grammatical or semantic rules to handle (2b-c) and constrain combinatorial possibilities (Blahuš 2009), plus a lexicon specifying as many known simplex words and compound frequencies as possible, because statistics show that if there *is* a simplex reading, this will almost always be the intended reading (1a-b), and in an x+y+z cut with (y+z) being a known, frequent compound, analysis x+(y+z) will almost always outrank (x+y)+z (3a-b).

3. Building a morphological lexicon

In order to identify candidates for non-simplex words, we extracted lemma lists from both EspGram's parser lexicon (49,700 non-name words) and a large monolingual Esperanto dictionary, PIV[3] (Plena Ilustrita Vortaro, Duc Goninaz 2005) with 42,800 non-name words, both together covering 51,500 unique non-name words. Next, we built a dictionary of simplex words (single roots without affixation) from three different sources: (1) Berlina Komentario (Pabst 2014), (2) the teaching website edukado.net[4] and (3) root-marked lemmas from PIV[5]. Together, these amounted to 17,100 unique roots.

We then inactivated ordinary lexicon lookup in the tagger, but allowed its morphological analyzer stage access to both the lexicon and our new root list. This way, the analyzer was forced to treat all words as "unknown", trying to assign compound- and affixation-based heuristic analyses wherever possible. Confronted with our lemma lists as line-separated text input, the system would identify a subset of potential compounds and affix-containing words, marking the rest as heuristic simplex words with an endings-based POS. This way 30,218 CW candidates were identified (58.7% of input lemmas) and submitted to manual post-editing of morpheme boundaries. The method produced about 10% false positive CW candidates[6].

For multi-part CWs the original EspGram analyzer would be content to provide a 2-part analysis even where one part was itself a CW, as long as the second part would permit the assignment of syntactic-semantic features from the lexicon. In order to achieve complete morphological decomposition, we reprogrammed the analyzer to look for combinations of root compounding and multiple prefixation/suffixation, providing patterns for acceptable combinations of the latter. In addition, we ran the process iteratively, using already-sanctioned CW analyses to break down CW parts that were themselves CWs. In the final CW annotation we distinguish between root boundaries (+), affix boundaries (%), inflexion ending boundaries (|) and ligature (~). Also, POS marks were added to first parts (e.g N: for nouns), to support semantic analysis or machine translation.

N:abel+reĝ%in|o (bee queen)
V:gard~o+tur|o (watch tower)
N:*har+sek%ig%il|o* (hair drier)
ADJ:*kaŝ+vojaĝ|ant|o* (stowaway)
ADJ:jun%ul%ar+gast%ej|o (youth hostel [young person group guest place])

The result is a CW lexicon with 28,849 CWs, amounting to over half (56.4%) of the overall non-name dictionary (cf. table).

4. Frequency distribution

In order to compare the prevalence of different CW types in the dictionary with token-based frequencies from running text, we ran the updated EspGram analyzer on a randomized Internet corpus (16 million words) from the Leipzig Wortschatz Corpora collection[7].

Almost half the dictionary lemmas (46.6%) were 2-part CWs. 5-part CWs were rare, but 3-part CWs still represented a respectable 9.1%. As running tokens, however, CWs are less frequent (19.4%), and only 2% of tokens had 3 or more parts. Participle derivation, in Esperanto a common source of noun/adjective creation (e.g. *fuĝ|into* [fugitive]), dropped by a similar margin, from 2% to 0.8%. However, most new/unknown tokens are CWs[8], and the effect is even bigger for multiple (3+) CWs, which were 10 times as frequent among new words.

[2] In the case of root compounding, an optional ligature-'o' may be appended to the first root for phonotactic reasons. This is marked with a tilde symbol ('~').

[3] http://www.eventoj.hu/steb/vortaroj/kapvortoj-piv/kapvortoj-en-piv2.htm (accessed 10/13/2015)

[4] accessed 11/7/2014: http://www.edukdo.net/instrumaterialoj?iid=11598&s=8f533f2f527e536d04e031d5be55c571

[5] information contained in www.eventoj.hu, but also Radikaro de Esperanto: http://mujweb.cz/malovec/rea.txt (10/13/2015)

[6] With a lemma list as input, this type ratio is of course different from a token ratio in running text.

[7] Specifically, the 2012 version of the 1M-sentence Esperanto corpus was used, from http://corpora2.informatik.uni-leipzig.de/download.html

[8] In fact, the real difference is even more pronounced, because most misspellings were heuristically counted as simplex words.

	dict lexemes		corpus (token%)	new in corpus (token%)
all lemmas	51205		15.5M	597K
CW lemmas	28860	56.4%	139658	110215
CW tokens			3005310	387609
simplex lemma	22345	43.6%	80.6%	33.42%
2-part	23856	46.6%	17.5%	44.5%
3-part	4670	9.1%	1.8%	19.1%
4-part	323	0.6%	0.2%	2.9%
5-part	10	0.02%	0.01%	0.13%
participles	1064	2.0%	0.8%	1.1%

Table 1: CW frequency breakdown

One interpretation for these numbers is that productive and systematic agglutination does play an important role in the Esperanto lexicon, and may increase transparency of (not least) new words for L2 learners, but on the other hand the effect appears to be less for the frequent words.

5. Performance evaluation

Because the whole parser lexicon was annotated and revised for compounds, and because the analyzer was tuned to try all possible compounding combinations for unknown words, the system's recall is almost 100%. Errors will therefore be related to over-generation of compound-splits (i.e. precision) and wrong morpheme boundaries. An obvious baseline for precision is the proportion of compounds that can simply be looked-up in the lexicon (87.1% of tokens or 21.1% of lemmas).

An inspection of 600 CW analyses for out-of-vocabulary word types, from 3 different frequency brackets (most frequent, least frequent and f=5), showed that false positives are rare (~ 1%) and limited to misrecognized upper-case items (names, e.g. *Silverman=sil+verm|an [silo worm-ADJ]*). All in all, 4.3% of new CW word types (6.4% of tokens) were misanalysed, missing morpheme boundaries being much rarer than wrongly placed or spurious boundaries.

600 new CW	types	tokens
all	4.3%	6.4%
wrong morpheme boundary	1.8%	1.5%
spurious morpheme boundary	1.2%	2.1%
missing morpheme boundaries	0.3%	0.7%
spelling error / foreign	1.0%	2.0%
of these: uppercase / name	0.8%	1.2%

Table 2: Performance on new CWs

6. A morpheme frequency dictionary

Once all words in running text are assigned a full morphological analysis, it is a fairly straight-forward task to build morpheme frequency lists. For the Leipzig web corpus (16M words), about 10,000 different recognized Esperanto morphemes occurred (i.e. 58% of our root lexicon). As expected, Esperanto's 42 affixes and 45 regular correlative pronouns had the highest ranks:

rank 1-10	-ig	rank 51-100	-ul, -ej, -an, ek-, -et kia/e, kiam
rank 11-20	-iĝ,mal-, -ad kiu, tiu	rank 101-150	-eg, -um,-em, -ism, iu, tion
rank 21-30	-ist, -on, tia/e	rank 151-200	-ind, -estr tiam, iom
rank 31-40	-ec, -aĵ, -ebl, -ar, kiel, ĉi	rank 201-250	-er, io
rank 41-50	-in	rank 251-300	ge- nenia/e, kial

Table 3: Affix ranks

However, these ranks are due to cumulative effect of many individual words, and a top-1000 frequency dictionary of Esperanto *words* (rather than morphemes) will, surprisingly, still only contain 25 root+root compounds and only 39 words with affixes. In our list, the respective firsts were, fittingly, *esper|ant%ist|o* (rank 267) and *inter+naci|a* (rank 309). This finding suggests that the agglutinative structure of Esperanto, and its effect on the lexicon learning curve, is important not so much for the absolute core dictionary, but rather as a passive reserve, and for lexicon expansion in intermediate learners.

7. Suffix-POS bigrams

While isolated morpheme frequencies are useful in a teaching context and for dictionary entry selection, it is also interesting - from a linguistic perspective - to shed light on the combinatorial properties of morphemes within an Esperanto word. Such information can be drawn from the same annotated data by focusing on "bigram" rather than "unigram" frequencies, and is useful, for instance, for assessing the likelihood of an unknown word to be correct in spell-checking and for semantically restricting morphological ambiguity (chapter 2). Leaving semantical root alternations to future work, we will here concentrate on the transition from suffixes to POS endings, i.e. on the likelihood which which a given suffix projects a certain word class (table 4).

Suffix	-o noun	-a (adj.)	-e (adv.)	-i/[aiou]s -[aio]n?t (verb)	further affixes
-aĉ	**56.7**	5.8	1.2	32.4	3.9 (ad,ul)
-ad	**86.7**	1.7	0.1	11.1	0.5

-aĵ	**97.0**	1.6	0.2	0.1	1.2 (ar,et,ej)
-an	**84.3**	8.3	0.1	0.1	7.2 (ism,in)
-ar	**81.1**	9.6	0.2	0.2	8.8 (an)
-ĉjo	**100**	-	-	-	-
-ebl	1.7	**68.2**	12.5	13.1	4.6 (ec,aĵ,ig)
-ec	**86.5**	10.3	1.6	1.0	0.1 (an)
-eg	31.4	**50.0**	1.7	13.8	3.0 (an,ul)
-ej	**95.9**	2.4	0.0	0.0	1.7 (an)
-em	27.4	**52.6**	10.6	1.6	7.8 (ul)
-er	**70.7**	9.7	3.2	10.1	6.3 (et,iĝ,ig)
-estr	**87.4**	8.9	0.0	0.2	3.5 (in,ar)
-et	**80.9**	5.3	0.3	11.5	2.1 (ad,aĵ)
-id	**61.3**	16.2	3.4	5.0	14.0 (in,et)
-ig	13.9	6.0	0.2	**76.2**	3.5 (ad,ebl)
-iĝ	20.9	0.4	0.0	**77.3**	1.4 (ad,em)
-ik	**90.6**	6.6	-	**-**	2.8 (ist)
-il	**88.6**	6.5	0.8	0.7	3.4 (ar,et,ist)
-in	**92.1**	4.1	0.1	0.1	3.6 (iĝ,et)
-ind	-	**65.3**	27.3	0.0	7.4 (aĵ)
-ing	**76.6**	9.9	0.3	7.3	5.9 (ebl,ej)
-ism	**65.6**	32.8	0.9	0.1	0.7 (an)
-ist	**80.1**	3.2	0.0	0.0	16.6 (ar,in)
-iv	32.5	**54.7**	4.5	2.9	5.3 (ec)
-iz	11.5	7.2	2.3	**47.8**	31.2 (ad,ig,ist)
-nj	**100**	-	-	-	-
-obl	0.9	**89.2**	7.3	-	2.6 (aĵ,ig)
-on	**58.5**	23.6	8.9	3.3	5.7 (ig,iĝ)
-op	21.2	**58.0**	16.5	0.8	3.55 (aĵ,ec,et)
-oz	35.7	**59.7**	2.3	1.2	1.2 (ad)
-uj	**90.8**	1.7	0.1	0.2	7.3 (ar,an)
-ul	**85.1**	0.2	0.1	0.1	14.5 (in,ar)
-um	30.9	6.7	11.9	**40.1**	10.4 (ad,il)

Table 4: Suffix/POS (token frequencies)

The highest affinity to one POS (noun) are found for the diminutive name suffixes (male -ĉj and femaile -nj), places (-ej), person suffixes (-ul, -in, -estr, -an) and thing suffixes -aĵ [thing], -il [tool] and -uj [container]. There seems to be most POS variation in the abstract suffixes, such as transitivity markers (-ig, -igx), -eg [intensity/size], -em [liking/propensity], -aĉ [peiorative] and the vague "default" suffix -um [associative]. While almost all suffixes allow a transition to a further suffix, some are more likely to do so, in particular -iz [supplying with], -id [offspring] and the person suffixes -ist [profession] and -ul [characterized by].

8. Conclusion

We have presented and evaluated a new lexicon of 28.800 Esperanto CW lemmas, and an improved version of a wide-coverage morphological analyzer with an accuracy of 93.6% for out-of-vocabulary CW tokens. These resources were used to build a frequency dictionary of 10.000 Esperanto morphemes, intended primarily for pedagogical purposes. CW frequencies appear to confirm the assumed high modularity and transparency of the Esperanto lexicon, but further research - not least a comparison with similar morpheme dictionaries for other languages - is needed to corroborate this typological claim. In addition, the CW lexicon and/or analyser can be used to examine semantic links between morphemes in a statistical fashion, such as the correlation between affixes and part of speech.

9. Bibliographical references

Bick, Eckhard (2007). Tagging and Parsing an Artificial Language: An Annotated Web-Corpus of Esperanto, In: *Proceedings of Corpus Linguistics 2007, Birmingham, UK*. (http://ucrel.lancs.ac.uk/publications/CL2007/)

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.

Blahuš, Marek (2009). Morphology-Aware Spell-Checking Dictionary for Esperanto. In: Petr Sojka & Aleš Horák (eds.): Proceedings of RASLAN 2009

Duc Goninaz, Michel [ed.] (2005). Plena Ilustrita Vortaro (PIV). Revised edition of Waringhien [ed.] (1970).

Hana, Jiří (1998). Two-Level Morphology of Esperanto. Master Thesis at MFF UK Praha, Praha. http://www.ling.ohio-state.edu/~hana/esr/thesis.pdf

Pabst, Bernhard (2014). Berlina Komentario pri la Fundamento de Esperanto - Dua Parto, Vortaro Oficiala. 5a eldono. (6/20/2014: http://esperanto-akademio.wikispaces.com/Berlina+Komentario+pri+la+Fundamento+de+Esperanto)

Quasthoff, Uwe; Fiedler, Sabine; Hallsteinsdóttir, Erla (eds.). 2014. Frequency Dictionary Esperanto. Leipziger Universitätsverlag.

Tellier, Angela and Roehr-Brackin, Karen (2013). The Development of Language Learning Aptitude and Metalinguistic Awareness in Primary-School Children: A Classroom Study. Discussion Paper. Essex Research Reports in Linguistics, University of Essex, Colchester, UK. [http://repository.essex.ac.uk/5983/1/errl62-1.pdf]