Tagging and Parsing an Artificial Language
An Annotated Web-Corpus of Esperanto

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

Summary

This paper presents and evaluates EspGram - a Constraint Grammar (CG) -based parser for the artificial language Esperanto. The parser was used to annotate a newly compiled, web-searchable corpus (18.5 million words), and achieved accuracy rates (F-Scores) of 99.5% for part of speech and 92.1% for syntactic function/dependency.

1. Introduction

In the first half of this paper, we present and evaluate EspGram - a Constraint Grammar (CG) -based parser for the artificial language Esperanto. The second half of the paper describes the compilation and annotation of a corpus of 18.5 million words covering Esperanto literature, news text and web pages.

As a planned language, conceived to be easy to learn and flexible to use, Esperanto has a highly regular morphology, where clearly perceived morphemes match linguistic categories almost one-on-one. Also, the core lexicon of the language was designed to avoid unnecessary ambiguity. Thus, morphological/lexematic ambiguity is almost entirely restricted to cross-compound ambiguity, and the average number of morphological readings is 1.12 readings per non-name word, as opposed to around 2.0 for most natural languages (depending on the way ambiguity is counted). Though since its inception (Zamenhof, 1887), the language has been allowed to evolve as a living system, most changes have occurred at the lexical level, and the morphological system remains largely unchanged. On the other hand, the relatively free word order of the language in combination with syntactic usage influence from different natural languages[1] has led to a language system and a speaker community very tolerant of syntactic variation, where norms are statistical rather than absolute.

This situation has important bearings on both parsing technology and corpus linguistics. First, with a reduced need for disambiguation, a part-of-speech tagger can be assumed to be almost identical to a morphological analizer, while a syntactic parser will face a number of challenges. Second, a corpus of correct but international Esperanto may offer interesting insights in lexical and syntactic variation, reminiscent of the variation of non-native, international English, the difference being that in Esperanto, such variation is not stigmatized, but rather allowed or even supported by the flexibility of the language system.

2. The parsing system

2.1 The morphological analyser

Like other Constraint grammar[2] systems (Karlsson et al., 1995), EspGram is a rule based system, applying contextual rules to handle morphological disambiguation and syntactic analysis. Input to the rule system is provided by (a) an NLP lexicon and (b) a morphological analyzer. In principle, the latter can achieve full morphological tagging simply by cutting an Esperanto word into

[1] Even native speakers of Esperanto are bilingual, having grown up in a multilingual environment.

[2] For an overview of different language systems, *cf.* http://beta.visl.sdu.dk/constraint_grammar.html

morphemes. Major word classes and tenses are marked by vowels, while number, case and verbal finity are marked by consonants:

-o (noun), -a (adjective), -e (adverb), -i (infinitive/base verb)
-j (plural), -n (accusative), -s (finite verb), -t- (passive participle), -nt- (active participle)
-as, -at-, -ant- (present tense), -is, -it-, -int- (past tense), -os, -ot-, -ont- (future tense)
-u (imperative), -us (conditional)

The word 'virin**o**jn' for instance, is analyzed as follows:

vir	in	o	j	n
'vir'	(female)	Noun	Plural	Accusative
[root]	[affix]	[PoS]	[number]	[case]

In the example, the lexeme base is *virino (woman)*, itself derived from the root *vir(o) (man)*. The language has a semantic system of prefixes and suffixes, as unambiguous and analytical as the grammatical endings system. The only possible ambiguity, then, arises where compounds clash with simplex words, affixed words or each other:

insekto *(insect)*
in/sekto *(feminist sect, used as a pun)*

Though any root can be made to change word class (*virina - womanly, ina - female, virine - in a womanly fashion)*, the vowel ending guarantees that Cross-PoS ambiguity cannot arise between major word classes, though it in theory can occur between function words and content words, since the former have no regular endings. The smallest possible tagging lexicon for Esperanto, then, is one that contains uninflectable function words ending in a vowel, '-j' *(cave* plural), '-n' (*cave* accusative) or '-s' *(cave* finite verb), *i.e.* words like *kaj* (and), *tri* (three), *kion* (what).

Such a tagger will not, however, be able to safely handle the semantic affixation system, making it impossible to pass a words semantic class or valency potential on to the syntactic module. In practice, therefore, a parsing lexicon is still needed for a good system. In the case of EspGram, a lexicon of 28.000 lexemes was built from the data base of a bilingual Esperanto-Danish dictionary (Bick) and a Danish-Esperanto machine translation (MT) system (http://beta.visl.sdu.dk/MT.html). The lexicon was then enriched with (a) so-called valency-potentiality tags and (b) semantic prototype markers. In the CG parsing paradigm, such tags are called *secondary tags*. Secondary tags will not be disambiguated themselves, but provide valuable context for the disambiguation of the primary tags (part of speech, inflexion, syntactic function and dependency).

Valency and semantic information can be collected in different ways:

- traditional manual lexicography
- corpus based studies
- morphological clues

Good mono- or bilingual dictionaries, especially learner's dictionaries, will list valency informations, such as transitivity, but only for verbs, and in the semantic area will list domain type, but leave semantic ontological classification to specialised works like wordnets. In our case, we completed the missing semantic information by lexical transfer from the Danish MT lexicon which did feature a full ontology. Corpus studies were used to fill in missing valency information from raw text in a bootstrapping manner (e.g. N PRP and ADJ PRP bigrams with mutual information), but should be repeated with the annotated corpus at a later stage.

The third method, morphological clues, will for most languages only work for a few specific

cases, like for the affix '-ist' as an identifier for the semantic class of human professional or "ideologist", and for transitivity are unsafe at best (e.g. '-ize'/'-ise'). In Esperanto, however, affixes do have a save meaning and provide useful valency and semantic classes:

-ig (<vt> transitivity): *kolorigi (to colour), sanigi (to cure), lumigi (to light up)*

-iĝ (<vi> intransitivity): *malsaniĝi (fall ill), prezidentiĝi (become president)*

-ul[3] (<H> human, person): *krimulo (criminal), saĝulo (wise man)*

-in (female), -id (offspring), ge- (couple), -ist (professional) ...

-uj (<con> container), -ej (<top> place), -aĵ (<food>), -il (<tool>) ....

While the traditional lexicon still had to be constructed for simplex (root-) words, affix classification covered a lot of what would have been manual lexicography in other languages. Currently, the root lexicon contains 28.000 semantically classified lexemes, which in turn support the affix method by sanctioning root candidates, thereby reducing ambiguity between affixed and simplex readings[4].

The table below shows, for a number of classes, the percentage of tokens in running text that can be classified by affix alone.

valency or semantic category (tokens / types)	affix	affix token count	affix-marked token %	affix type count	affix-marked token %
<vt> transitivity (50,358 / 12,160)	-ig	4,659	9.3 %	1,856	15.3 %
<vi/ve> intransitivity (34,889 / 8,675)	-iĝ	4,252	12.2 %	1,495	17,2 %
<ve> ergative (9,688 / 1,649)	-iĝ	1.584	16.4 %	825	50,0 %
<con> container (1,192 / 126)	-uj	66	5.5 %	27	21.4 %
<L...> <inst> place (12,400 / 1,806)	-ej	1,614	13.0 %	344	19.0 %
<tool> tool (713 / 179)	-il	345	48.4 %	99	55.3 %
<H> <Hprof> <Hfam> ... human	-ul	1,673	8.6 %	522	17.3 %
(19,503 / 3,017)	-in	1,311	6.7 %	237	7.9 %
(not counting human groups <HH>)	-id	27	0.1 %	8	0.3 %
	-ist	2,941	15.1 %	621	20.6 %
	-an	845	4.3 %	196	6.5 %
	-estr	294	1.5 %	55	1.8 %
	-nj, -ĉj	70 (62, 8)	0.4 %	7 (3,4)	0.2 %
	ge-, bo-	79 (74, 5)	0.4 %	0	0.0 %
	-ant, -int, -ont	3,158	16.2 %	719	23.8 %
	all human	10,398	53.3 %	2,365	78.4

Table 1: affix-based determination of lexical category

[3] Theoretically, '-ul' can be used for to other semantic types, trees <Btree> and ships <Vwater>. These cases are not, however, productive in modern language, and with all older forms listed in the lexicon, a parser can safely assume the affix to be unambiguous.

[4] The ending '-nto', for instance, denoting present participle nouns, is safe with a verbal root (e.g. *falanto - speaker),* but does occur in simplex words like *ganto - glove* or *kanto - song* and their compounds.

As can be seen from the table, the affix "hit rate" is generally higher for types than for tokens, reflecting the affixes productive nature and the un-affixed core-vocabulary's frequency. The affix rate was highest for the group of human prototypes (around 53.3% for tokens and 78.4 for types) as well as tools , while it was low for containers, with a considerable token-type difference (5% vs. 21.4%), probably reflecting the fact that containers are (a) not a very productive class, and (b) largely covered by frequent simplex words such as *taso (cup), sako (bag) etc*.

For the valency category of transitivity, ergatives have a higher affix ratio than transitives, a difference particularly marked at the type level, possibly because of the productive inclusion of noun roots *(become s.th.)* in the former.

All in all it is obvious that a large part of the lexicon in running text can be class-typed based on affixation rather than traditional word nets or valency dictionaries. Apart from affixation, the main lexicon-bootstrapping method employed was pattern extraction from large corpora iteratively annotated with increasingly accurate versions of the parser.

2.2 The syntactic parser

The disambiguation an syntactic rules i EspGram are formulated in the Constraint Grammar fashion, removing, selecting or mapping token-based category tags, based on sentence-wide context conditions. Systematic use was made of morphological category markers, semantic affixes, domain markers and valency information.

All in all, the grammar contains 1,498 rules, with the following breakdown:

Morphological/PoS section
- 51 REMOVE rules
- 21 SELECT rules

Syntactic section
- 644 MAP rules (+ 29 ADD rules)
- 541 REMOVE rules
- 212 SELECT rules

While CG's for other languages typically invest more rules in the PoS/morphology section than in the syntax section, the percentage of the former is only 4.8% in EspGram. Even those few morphological rules that do come into play, are largely "syntactic" in their nature, reflecting design choices as to where (on which linguistic level) to express a given ambiguity. Thus, certain subordinators *(kiel, kio, kion, kiu ...)* are disambiguated as either relative <rel> or interrogative <interr>, and a number of prepositions is tagged as adverbs when used to pre-quantify numbers :

ĉirkaŭ kvincent dolaroj - *about* 500 dollar

ĝis 15 partoprenantoj - *up to* 15 participants

The only real part of speech ambiguity is between proper nouns and other word classes in sentence initial position. Names have a notoriously unstable orthography in Esperanto, with three systems used in parallel:

(a) fully translated names (countries, major towns). These names feature the obligatory -o noun ending and will take the -n marker in the accusative case: *Danio (Denmark), Gronlando (Greenland), Munkeno (Munich)*

(b) phonetically adapted names, exploiting the phonetic regularity of the Esperanto alphabet for a kind of transliteration: *Buŝ (Bush), Ĥruŝĉov'o (Khrushchev)*, with or without Esperanto endings.

(c) "raw" names, taken literatim from source languages with a Latin alphabet (though possibly with loss of or changes in diacritics)

Across these conventions, names ending in 'on', such as the author *Claude Piron* or the politician *Clinton,* can be case-confused with accusative forms of hypothetical *Piro*[5] or *Clinto,* if they are not int the system's lexicon. Here, CG disambiguation will use contextual clues for disambiguation, for instance:

REMOVE (ACC) (-1C PRP) (NOT -1 PRP-DIR OR PRP-LOC) ; *remove the accusative reading (ACC), if there is an unambiguous (C) preposition (PRP) at the -1 (i.e. immediately left) position, unless (NOT) this preposition is directive or locative - in which case it might govern a direction-accusative in a place name.*

The syntactic level of the EspGram grammar consists of (a) a mapping level, assigning potential syntactic functions according to word classes and immediate context, and (b) several layers of full-context disambiguation rules which remove or select these mapped function candidates until only one survives per token. Rule layers are applied iteratively with the last layers containing the most heuristic (i.e. least safe) rules.

A syntactic tag can consist of two parts - the function itself and a dependency direction marker. @SUBJ> and @<ACC, for instance, mark a subject and object positioned, respectively, left and right of their verbal heads. A dependency marker may be specified as to what it attaches to. Thus, @N< is a postnominal dependent, @P< the argument of a preposition, with the N and P denoting the PoS type of the head.

The following is an example of a syntactic disambiguation rule:

REMOVE (@<SUBJ) (*-1C @ARG/ADVL> BARRIER VFIN)

This rule weeds out crossing attachment brackets at the clause level, removing left attaching subjects if there is a safe (C) *right*-attaching argument or adverbial anywhere (*) to the left (-1) without a finite verb (VFIN) in between.

Since every token is assigned a dependency marker, and subclause function is marked on subordinated verbs, the CG annotation can encode a complete syntactic tree, albeit with a certain degree of underspecification: A postnominal attachment marker on a preposition, for instance, rules out ad-verbal pp-attachment, but does not specify the attachment order of *multiple* postnominal pp's.

A full syntactic tree can be constructed in two ways:

(a) adding a phrase structure grammar layer with a PSG rules operating on CG function tags rather than terminals as the smallest unit of structure, e.g.

STA:fcl = SUBJ> P (<ACC, <ADVL, <SC)*
X:np = >N X:n N<

where the first rule will mount a finite clause from a subject, predicator and optional other constituents, and the second will assemble an np from a noun-head and pre- and postnominals, while raising head word function (X) into np function.

(b) adding an attachment grammar and dependency rules specifying unambiguous dependency arcs between a CG daughter *function* and a head *form*, e.g.

@FS-N< -> (¤NPHEAD) IF (L) TRANS:(<rel>) BARRIER:(PR,IMPF,<co-fin>)
@>A -> (ADJ,ADV,DET,NUM,PCP1,STA) IF (R) NOTHEAD=(<aquant>.*@>A)

where the first rule attaches a relative clause to a token carrying a np-head function after looking

[5] *Piro* is particularly tricky, since it also is a name meaning 'pear'

left (L) across (TRANS) a relative pronoun (<rel>) if one can be found without an interfering (BARRIER) finite verb (PR,IMPF) or finite verb coordinator (<co-fin>). The second rule attaches a pre-adjectival or pre-adverbial modifier to a token of the right wordclass to the right (R), but exempts intensifiers that are themselves premodifiers in the same phrase.

Both methods (a) and (b) were implemented in EspGram, and VISL filters (http://beta.visl.sdu.dk/treebanks.html) are available for converting PSG and dependency formats into each other. However, research on other languages (Bick, 2005-2) suggests a considerably higher efficiency and structural recall to the dependency method (b), which appears to be more robust in the face of function tag errors in the input, and will construct more complete trees than the PSG method (a), even when compared in the format of the latter.

Source code examples of the two annotation styles are given in table 2. Though (1) ellipsis, (2) coordination and (3) discontinuities may introduce complications for either the dependency (1-2) or the constituent format (3), the two are roughly information equivalent[6]. Thus, the numbered dependency markers in notation (a), e.g. #9->6 (word 9 attaching to word 6) allow the definition of head-driven constituents (b), where bracketing depth is shown as =-indentations.

(a) Constraint Grammar Dependency notation	**(b) VISL Constituent Tree notation (PSG)**
En la tria grupo *(In the third group)* kuniĝis *(came together)* tiuj, kiuj malaprobas *(those who criticize)* *ĉ*ion, kio okazis *(all that happened)* en la katolika eklezio *(in the Catholic Church)* dum la pasintaj dudek jaroj *(during the past twenty years).*	
	STA:fcl
	=fA:pp
En [en] PRP @ADVL> #1->5	==H:prp("en") En
	==DP:np
la [la] ART @>N #2->4	===DN:art("la") la
tria [tria] <num-ord> ADJ S NOM @>N #3->4	===DN:adj("tria" <num-ord> S NOM) tria
grupo [grupo] N S NOM @P< #4->1	===H:n("grupo" S NOM) grupo
kuniĝis [kunigxi] <mv> V IMPF @FS-STA #5->0	=P:v-fin("kunigxi" <mv> IMPF VFIN) kuniĝis
	=S:np
tiuj [tiu] <dem> PRON DET P NOM @<SUBJ #6->5	==H:pron-dem("tiu" <dem> DET P NOM) tiuj
$, #7->0	==,
	==DN:fcl
kiuj [kiu] <rel> PRON DET P NOM @SUBJ> #8->9	===S:pron-rel("kiu" <rel> DET P NOM) kiuj
malaprobas [malaprobi] <mv> V PR @FS-N< #9->6	===P:v-fin("malaprobi" <mv> <np-close> PR VFIN) malaprobas
	===Od:np
ĉion [cxio] <quant> PRON INDP S ACC @<ACC #10->9	====H:pron("cxio" <quant> INDP S ACC) ĉion
$, #11->0	====,
	====DN:fcl
kio [kio] <rel> PRON INDP S NOM @SUBJ> #12->13	=====S:pron-rel("kio" <rel> INDP S NOM) kio
okazis [okazi] <mv> <np-close> V IMPF @FS-N< #13->10	=====P:v-fin("okazi" <mv> <np-close> IMPF VFIN) okazis
	=====fA:pp
en [en] PRP @<ADVL #14->13	======H:prp("en") en
	======DP:np
la [la] ART @>N #15->17	=======DN:art("la") la
katolika [katolika] ADJ S NOM @>N #16->17	=======DN:adj("katolika" S NOM) katolika
eklezio [eklezio] N S NOM @P< #17->14	=======H:n("eklezio" S NOM) eklezio
	=====fA:pp
dum [dum] PRP @<ADVL #18->13	======H:prp("dum") dum
	======DP:np
la [la] ART @>N #19->22	=======DN:art("la") la
pasintaj [pasi] V PCP AKT IMPF ADJ P NOM @>N #20->22	=======DN:v-pcp("pasi" AKT IMPF ADJ P NOM) pasintaj
dudek [dudek] <card> NUM P @>N #21->22	=======DN:num("dudek" <card> P) dudek
jaroj [jaro] <clb-end> N P NOM @P< #22->18	=======H:n("jaro" <clb-end> P NOM) jaroj

Table 2: Dependency vs. PSG analysis

[6] For more information, *cf.* http://beta.visl.sdu.dk/treebanks.html

3. Evaluation

The performance of EspGram was measured against a hand-annotated gold standard corpus of news text produced in Esperanto by contributors embedded in a variety of cultures and matrix language commuities. The test chunk, from the *Monato* magazine, contained 4,400 tokens (3,439 function-carrying words). On these data, current parser accuracy rates (F-scores) run at 99.5% for part of speech and 92.1% for syntactic/dependency.

	Recall	Precision	F-score
Base form / lexeme	99.7	99.7	99.7
PoS (part of speech, word class)	99.5	99.5	99.5
Morphology / inflexion	99.7	99.7	99.7
Syntactic function	93.4	90.9	92.1

Table 3: Parser performance

While encouraging, in a cross-language comparison, these numbers confirm the hypothesis that Esperanto is easier to tag (morphologically) than to parse (structurally), and poses a syntactic challenge on par with other languages. Thus, even with the small system presented here, the PoS and morphological error rates (0.5 and 0.3%, respectively) were even lower than the already excellent PoS error rates of comparable CG systems for Danish (1.3%, *cf.* Bick 2003) or Spanish (1%, *cf.* Bick 2006), while the syntactic error rate (8%) was higher than in similar CG systems for Danish (5%, *cf.* Bick, 2003) and Spanish (4.7%, Bick, 2006), and in terms of recall also higher than in Lingsoft's English ENGCG (Lingsoft, 2007) and the Estonian CG described in (Müürisep and Uibo, 2005), though the good recall of the latter (97-98% and 98.5%, respectively) must be seen in the light of a somewhat lower precision (85-90% and 87.5%, respectively). It must also be born in mind that all CG systems, notwithstanding they mutual differences, compare favourably with probabilistic approaches. Thus, the best performing dependency parsers in this years machine learning shared task at the ConLL conference achieved a syntactic label accuracies between 80.9% (Basque) and 93.1% (English), even with *manually corrected* PoS input (Nivre et al., 2007).

Similarly good results for Esperanto PoS/morphology were reported by Warin (2004), who compared his own rule-based system (PDP11, 99.3% correct PoS/morphology) with a stochastic tagger (TnT, 98.6% average accuracy). Since even the stochastic tagger performed better than usual for other languages, it is reasonable to assume that part of the accuracy gain was due to the specific - and regular - traits of Esperanto morphology.

In the syntactic field[7], on the other hand, Esperanto is not only a challenge because of a freer word order than found in Danish or Spanish, but also because its international speaker community is liable to exploit a large portion of its structural possibilities under the influence of different native languages. A qualitative error analysis of the test corpus thus demonstrated some syntactic variation likely to be caused by first language interference.

For instance, speakers of Slavic languages have a tendency to omit the definite article before "name-like" nouns in Esperanto, and in general do not always follow conventions established by Germanic and Romance Esperanto speakers:

[7] No other syntactic Esperanto parser was available for comparison. (Lin and Sung, 2004) used apartial parsing with a Transformation-Based Learner system PSG, but because of complexity issues only discuss sentences with 3-5 words, where 1 out of 30 sentences was "correctly" parsed in the following sense: *Since merely the path with lowest/best score are considered right, and we have no external data to decide if some higher scored rule should be the right one, we can just demonstrate the distance between our result and the ideal* (quote from chapter 2.1)

La? speciala komisiono de **[la]** sovetia registaro en 1943 venis al **la** alia konkludo.

(The /A special committee of the Soviet government in 1943 arrived at another conclusion.)

Article usage in the example is not counter to any formal rules, but the statistical norm would omit the third article (possibly the firs) and insert the second.

Another non-standard variation is the complementation and placement of participles sometimes used by Slavic and Japanese speakers:

Tiel la filmo estas duoble uinda de esperantistoj.

(Thus, the film is doubly enjoyable by esperantists.)

Nun jam planite estas, eksporti la filmon al la tuta mondo.

(Now already planned is to export the film to the whole world.)

In the first example the adjectivally suffixed form 'gxuinda' (enjoyable) is both intensifier-modified like an adjective (doubly), but at the same time carries an agent pp (by esperantists) as in a participle clause. In the second example, the participle *planite* (planned) is placed before the copula verb rather than after, as would be statistically more normal. While such variation does not hinder human understanding of the sentence and, in fact, is part of the creative potential of the language, it makes it more difficult for at parser to establish correct constituent borders and attachments.

When the syntactically analysed test chunk was used to construct full tree structures, the dependency method proved not only, as predicted, more robust than the psg method, but also considerably faster:

200 sentences average sentence length 17 words	PSG-method raw CG input / revised CG input	Dependency method raw CG input / revised CG input
attachment accuracy[8]	-	88.9 % / 97.7 %
partial/malformed trees	53.5 % / 50.5 %	-
trees with circularity warning	-	2 / 1
system time	44.1 sec / 40.3 sec	0.046 sec / 0.040 sec
user time	104.6 sec / 95.8 sec	11.6 sec / 11.5 sec

Table 4: Comparison of PSG and dependency tree generators

Since most syntactic function errors will cause at least cause one attachment error, the dependency trees had an attachment accuracy several percentage points beneath the function tag F-score. However, on corrected CG input, attachment accuracy rose to 97.7%. One methodological difference between the psg and dependency methods was that when the output of the latter was transformed into constituent trees, even wrong trees would be mostly well-formed (since only 1 or 2 trees had formal dependency defects in the form of circularities), while about half of the psg-generated trees were incomplete, i.e. parse failures with only partial structures. From a corpus or treebanking perspective, this inherent difference in the percentage of well-formed trees can be seen as a further advantage of the dependency method, since well-formed trees are more accessible to treebank manipulation and search tools.

[8] while the other figures in the table were calculated for 200 sentences, attachment accuracy was only evaluated in a quarter of these, amounting to 956 words.

4. The Esperanto on-line corpus

4.1. Corpus creation

With its small diaspora language community without big financial or cultural institutions, let alone a tax or governmet base, esperanto is, in socio-linguistic terms, a minority language, and the limited amount of language technology available reflects this. Thus, when our project was conceived in 2003/4, only one corpus project (*Tekstaro de Esperanto*[9], *cf.* ESF, 2005) existed, and although following the EU's Text Encoding Initiative (TEI), it did not address grammatical annotation. However, the Esperanto community does produce a relatively large amount of written text in the form of magazines, books and, not least, easily available internet based material, such as Wikipedia articles.

From these sources, we compiled a corpus of about 18.5 million words[10], consisting of both traditional files - such as newspaper back issues - and material acquired with a web crawler[11]. The distribution of the corpus is about 50% literature (including some classical Zamenhof texts[12]), 17% news text (mostly the Newsweek-style international magazine *Monato,* and the more Esperanto-centered *Eventoj)*, 17% Wikipedia[13], as well as 16% mixed web pages and personal e-mail:

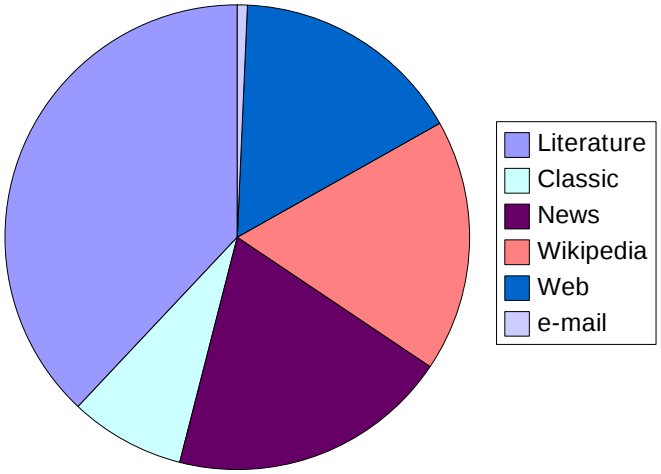

Figure 1: Distribution of corpus sources

In order to turn the collected data into a true corpus, we cleaned the texts of binary data, html and other meta data, and a preprocessor assigned sentence separation marks and chunk id's. Also encoding schemes needed attention, harmonizing material in iso-latin, utf8 *etc.*, because Esperanto features 5 non-standard accented letters in its alphabet (the consonants 'ĉ', 'ĝ', 'ĥ', 'ĵ', 'ŝ' with a circumflex, and the semivowel 'ŭ'. These letters are not part of the iso-latin-1 set, and encoded in a number of different ways, among them html codes and h- and x-conventions, replacing the accent with an added 'h' (classical style) or 'x' (alternative modern style): *charma = cxarma = ĉarma (*English: *charming).*

[9] http://bertilow.com/tekstaro/

[10] The complete collection of internet texts is considerably larger, but favouring a "clean" core corpus, we have not yet used all available data.

[11] The web crawler was programmed in 2004 by Jacob Nordfalk as part of a joint project aimed at building a text and tool base for esperanto lexicography and mobile phone applications.

[12] These include *La Biblio (The Bible)* and *Fabeloj (H.C. Anderson).*

[13] This consitutes all of the 2004 Esperanto Wikipedia. The current Wikipedia database for the language is about 7-8 times bigger, and this section of the corpus is clearly a candidate for yearly updates.

A special program, *esponly*, was written to filter out non-Esperanto text, which was present at the document level in the e-mail section and the sub-document level in the web section. *Esponly* works at one line at a time and assigns language scores, based on typical letter combinations and key words, for both Esperanto-like text traits and English, German, French *etc* traits. A line is accepted as being in Esperanto, if three conditions are fulfilled:

1. the Esperanto trait count is higher than the sum of foreign-language trait scores
2. the Esperanto trait count is above a certain threshold
3. the foreign-language sum count is lower than a certain threshold

In order to avoid erroneous inclusion or exclusion of short lines or 1-word lists, a base value from the preceding trait scores is passed on to the next line. Thus the fate of short expressions will be decided of their left hand language context.

As a next step, the corpus was annotated with the EspGram system in consecutive tagging and parsing steps, and encoded in the CQP format of the IMS Corpus Workbench (Christ, 1994) for use in a graphical, freely accessible search interface (*CorpusEye*, http://corp.hum.sdu.dk). All texts are searchable for text, PoS and syntactic function, returning concordances and statistical overviews. All search categories and quantifier patterns can be "mounted" using menu-based choices.

Finally, a small part of the data was converted into a full-depth treebank, using a rule-based dependency grammar. The treebank is available in both the dependency and constituent tree formats.

4.2. Corpus uses: The example of genre-dependent lexical variety

If a corpus is to come anyway near a true reflection of an entire language system, it has to be genre-balanced across different sources (for practical reasons this will often mean "across *written* language sources). Also, certain text sources are important for a balanced corpus, because they can be said to contain a certain balance by themselves - thus news text has a good topic spread, while encyclopedic material (Wikipedia) guarantees a good lexical or even terminological coverage.

In our search interface (CorpusEye), we offer contrastive statistics on the different sections of the on-line corpus, comparing for instance the lexicon and syntax of classical text and modern news text, respectively. That lexical coverage varies a great deal, can be seen from Ill. 2, where the left columns show the absolute number of lexeme types (in thousands), and the right columns express lexeme variation (lexeme types divided by the square root of subcorpus size).

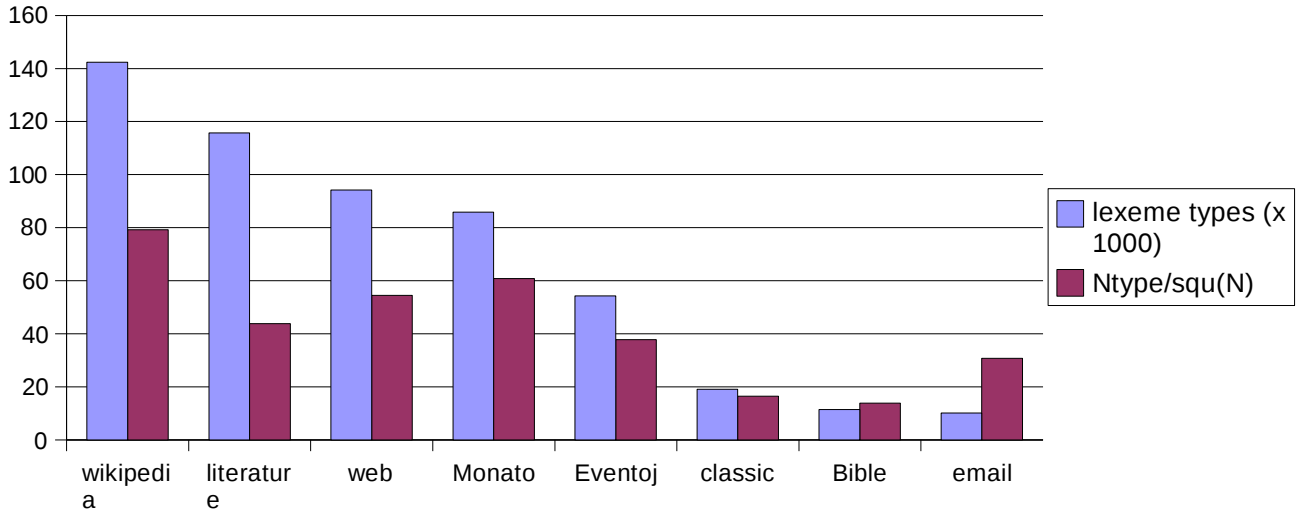

Illustration 2: lexical distribution

It becomes clear from the overview that classical Esperanto makes do with a much smaller lexicon, also in relative terms, while modern texts employ a 2-5 times larger lexicon, with the encyclopaedic genre of Wikipedia leading the pack[14].

Apart from lexicographical research, the search interface lends itself to syntactical studies, where more complex, category-based searches may be necessary. In the example concordance below, the search command asked for two adjacent fields, with a participle functioning as prenominal clause head in the first, and a preposition in the second:

Kaj veninte li/sxi devas estimi la **ekzistantajn en** tiu lando tradiciojn , morojn ktp .
(appreciate the existing-in-this-country traditions ...)
kaj evoluigas ellaboritajn de Lenin'o kaj **ekzamenitajn de** la praktiko bolsxevismajn principojn de la partia konstruado
(develop designed-by-Lenin-and-proven-by-practice bolshevik principles of party building ..)
2 . La Kongreso plene aprobas la **ellaboritan de** Centra Komitato de KPSU'o koncepton pri la akceligo
(... fully supports the worked-out-by-the-Central-Committee-of-[the]-KPDSU concepts about acceleration)
La Kongreso alte taksas **aplikatajn far** Centra Komitato rimedojn por perfektigo de la politika
(.. highly appreciates applied-by-[the]-Central-Committee means for a perfection of the political ....)
, ke la regantaj imperiismaj rondoj ne vidas la **implicitan en** tio minacon .
(... does not see the involved-in-this threat ...)
la principe alia kolektivisma bazo , kondukis la **irantajn laux** tiu vojo landojn al pli altaj niveloj de evoluo .
(.. leads the walking-along-this-path countries to higher levels of evolution ...)
- li diris - Vizitu lin kaj prenu de li **donacitan al** mi bokaleton da mielo .
(... Visit him and take by-him-offered-to-me little cup [?] of honning)

Pending further research, the examples present possible (i.e. not grammatically wrong), but rare constructions found mostly in the bureaucratic-ideological tradition of the adminsitration of former communist countries. Even the last example, not political, carries a Slavic language mark, the use of *bokalo* (a pharmaceutical term) instead of *pokalo* (a kind of cup), and thus also point toward Eastern European usage.

5. Conlusion and outlook

With a PoS accuracy of 99.5 percent and a syntactic function accuracy of 92%, the CG parser for Esperanto described in this paper has proven to perform roughly on par with CG systems for other languages, but results also reflect the specific traits of the Esperanto language system, with an easier-than-average morphology and a comparatively more difficult syntax. Given the fact that the parser was developed virtually without outside funding, and seeded with existing CG rules from other systems, it should be possible to achieve measurable improvements in syntactic performance in the future. This would also further facilitate the already promising conversion into dependency tree structures (97.7% correct attachment on correctly tagged input), and we intend to use EspGram to build a sizeable treebank for Esperanto in the near future. Both the treebank and the existing CG-tagged corpus should facilitate research into Esperanto lexicography, syntax and usage variation. To further improve the corpus to meet these ends, spelling variation should be quantified and news and web or e-mail sources typed according to the likely matrix language of their respective authors. Among others, two corpus research areas, already touched upon in this paper, should be further examined: (a) the varying usage of participle clauses and (b) lexical growth mechanisms in Esperanto. Thus, we hope to use our corpus to clarify to what extent new lexical material consists of

[14] For a more thorough evaluation, it might be preferable to use corpus chunks of identical size rather than the square root approximation. Also, the possible prevalence of spelling errors in the web data should be taken into account.

either loan words, or newly coined expressions employing the language's own semantic affixation and compounding system.

References

1. Bick, Eckhard. (2003) A CG & PSG Hybrid Approach to Automatic Corpus Annotation, In: Kiril Simow & Petya Osenova (eds.), Proceedings of SProLaC2003 (at Corpus Linguistics 2003, Lancaster), pp. 1-12

2. Bick, Eckhard. (2005-1) CorpusEye:Et brugervenligt web-interface for grammatisk opmærkede korpora, In: Peter Widell & Mette Kunøe (eds.), 10. Møde om Udforskningen af Dansk Sprog 7.-8.okt.2004, Proceedings. pp.46-57, Århus University

3. Bick, Eckhard. (2005-2) Turning Constraint Grammar Data into Running Dependency Treebanks, In: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.), Proceedings of TLT 2005 (4th Workshop on Treebanks and Linguistic Theory, Barcelona, December 9th - 10th, 2005), pp.19-27

4. Bick, Eckhard. (2006) A Constraint Grammar-Based Parser for Spanish. In: Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology (Ribeirão Preto, October 27-28, 2006).

5. Bick, Eckhard. (1990-1997) Esperanto-Dansk. Århus: Mnemo.

6. Christ, Oli. (1994) A modular and flexible architecture for an integrated corpus query system. COMPLEX'94, Budapest.

7. Esperantic Studies Foundation. Tekstaro de Esperanto. Esperantic Studies, 16/2005 (ISSN 1084-9831). http://www.esperantic.org/es16web.pdf. http://bertilow.com/tekstaro/

8. Karlsson et al. (1995) Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text. Berlin: Mouton de Gruyter

9. Lingsoft. (2007) Components of ENGCG. Accessed at http://www2.lingsoft.fi/ doc/engcg/ intro/components.html, 30 June 2007

10. Müürisep, Kaili and Uibo, Heli. (2005) Shallow Parsing of Spoken Estonian Using Constraint Grammar. In: P.J.Henriksen & P.R.Skadhauge, Proceedings of NODALIDA-2005 special session on treebanking. Copenhagen Studies in Language #33/2006

11. Nivre, J., J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. (2007) The CoNLL 2007 shared task on dependency parsing. In Proc. of the CoNLL 2007 Shared Task. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

12. Warin, Martin. (2004) A Comparison between a Rule-Based and a Stochastic PoS-Tagger for Esperanto. C-uppsats i Allmän språkvetenskap med inriktning på datorlingvistik. Available on-line from http://ling16.ling.su.se: 8080/PubDB/ doc_repository/ warin2004comp.pdf

13. Lin, Zhemin & Sung, Li-May. (2004) Tiny Corpus Applications with Transformation-Based Error-Driven Learning: Evaluations of Automatic Grammar Induction and Partial Parsing of SaiSiyat. Proceedings of PACLING 18. Tokyo

14. Zamenhof, L. L. (1887) Dr. Esperanto. Warszawa: Zamenhof