Enriching Data-Oriented Parsing by blending morphology and syntax

a pilot using a constructed language

Andreas van Cranenburgh*

May 14, 2010

Cognitive Models of Language project, April 2010,
Master of Logic, University of Amsterdam

Abstract

Esperanto is a constructed language with a rich and regular morphology. It seems likely that taking its morphology into account when parsing syntax will improve accuracy. I will investigate the effects of considering morphological and phrase structure analysis as separate, autonomous steps, versus combining them into a single DOP model. I will assume a hierarchical representation for both syntax and morphology.

We describe a formal grammar that enumerates the word forms of Esperanto. Using a DOP model sequences of morphemes and tags can be analysed and assigned a hierarchical structure. The resulting DOP model can either be merged with a syntactic treebank into a combined DOP model, or mapped to the leaves of the parse trees produced by a syntactic model, to obtain tree structures with both phrasal and morphological constituents.

The resulting system will be applied to a small corpus of morphology and syntax. Evaluation with a larger syntactic treebank, as well as the induction of morphology tags from dictionaries remains to be done.[1]

*acranenb@science.uva.nl, 0440949

[1] **Acknowledgements**: I wish to thank the following people (in reverse chronological order): Federico Sangati for practical advice on DOP, Ken Miner for advice on morphology and suggesting the application of DOP to Esperanto, Eckhard Bick for the Monato treebank, Rens Bod for teaching me Data-Oriented Parsing, and last but not least Wim Jansen for teaching me Esperanto.

Contents

1 Introduction

1.1 Research questions

> Does integrating morphology with syntax improve parsing results for syntax?
>
> Are morphology and syntax autonomous? ie., is morphology opaque or transparent to syntax?

These possibilities correspond to a modularist (cf. Pinker 1994, Jackendoff 2003) vs. an interactionist approach (cf. MacWhinney 1987). Here modularism refers to the functionalist hypothesis of the autonomy of syntax from other levels, both the stronger claim of processing autonomy (full modularity, encapsulation), and its weaker form of representational autonomy. In effect the issue at stake here is the nature of the morphology-syntax interface. On the one hand there is the extreme of syntax seeing only a Part-of-Speech tag (and possibly the unanalyzed word or lemma as well if syntax is lexicalised), on the other hand there is the other extreme where morphology is literally a part of syntax that has been conveniently ignored in the majority of work in (computational) linguistics to date. Jackendoff's (2003) parallel architecture suggest a compromise (at least for generativists) where interfaces of different autonomous levels are possible (e.g., phonology-semantics to deal with focus effects).

> Are words the smallest units of syntax, or is it perhaps morphemes? A fully transparent morphology could imply that morphemes are the smallest units of syntax, rather than words as is customarily assumed.

Because this is a pilot project, I shall only attempt to answer the first question, but this may provide a hint as to the other questions. In addition the answer to the first question shall only concern Esperanto, which has been chosen for its rich and regular morphology. Recent work by Tsarfaty (2010) underscores the need for adapting formalisms to morphologically rich and relatively free word-order languages such as Hebrew and Arabic, but also Esperanto. Without morphology-aware parsers it does not seem possible to attain the current parse accuracy of languages such as English and Chinese.

1.2 Approach

1. Construct a corpus of sentences annotated with phrase structures, and a lexicon of words annotated with morphological structures. The assumption is that while syntax may use information in morphology, morphology does not need information from syntax, hence the possibility of constructing a morphological corpus independent of the text corpus; note that this amounts to assuming that for the purpose of constructing a corpus the morphology is context-free.
2. Divide the corpus into training and testing, train on the former with DOP1 or DOP*[2] (Zollman 2005)
3. Measure performance of syntax model, this will be the baseline
4. Morphology transparent to syntax: take treebank corpus, merge phrase structure trees with morphological analyses, construct a single DOP model
5. Morphology opaque to syntax: construct a DOP model for morphology, taking one word at a time, and a DOP model for syntax, producing phrase structure trees without morphology. Morphological structure and phrase structure can be parsed in parallel and independent of each other.

2 Background

2.1 Morphology

Most work in Computational Linguistics focuses exclusively on syntax; this is a form of syntactocentrism, a term coined in Generative Linguistics (Jackendoff 2003). This also goes for Data-Oriented Parsing (DOP), although excursions into semantics have been made. In this project I will go in the other direction and turn to the lower linguistic stratum of morphology. Most accounts of morphology in Computational Linguistics seem to present the structure of words as merely a sequence of morpheme-feature pairs (e.g., Jurafsky & Martin 2000), as parsed by a Finite State Transducer (cf., Schmid et al. 2004). During the course of this project I even came across a master thesis describing such a two-level morphology for Esperanto (Hana 1998).

[2] More on Data-Oriented Parsing further on

However, due to the complexity and potentially unlimited productivity of morphology in Esperanto such a representation will necessarily reveal only part of the structural information of words in Esperanto (more on this in the next section). Such an approach is to the representation of the present project what POS tagged sentences are to hierarchical phrase structure trees. Although the present project focuses on Esperanto, the method of adding morphology to DOP should generalize to other languages, also to languages such as English which display only a very limited amount of morphological productivity and hence exhibit only a subset of the derivational complexity in morphologically richer languages.

2.2 About Esperanto

Esperanto is a constructed language (also referred to as a planned language). The term "artificial language" that is sometimes employed is inappropriate, as its artificial design is only a point in time of its century long continuous usage and evolution. It is a spoken language with its own literature and culture, so while it may not be a "natural language" strictly speaking (Gobbo (2009) uses the term Quasi-Natural Language), it is certainly a human language that performs all the communicative and expressive functions of Ethnic languages, albeit mostly as a second language used by a diverse and scattered speech community. When Esperanto is referred to in the popular press as a "failed project" this refers to the ambitious pacifistic ideals of the language, not its maturation as a langauge.

Typologically Esperanto has the unique character of being a morphologically agglutinative and synthetic language, yet with a vocabulary largely based on Romance languages (apart from some German & Russian words, and schematic function words as well). Its word formation is extremely compositional (ie., complex words are semantically fully transparent); I would go as far as to contend that it is the most compositional spoken language in use today. Its syntax is schematic (designed) and allows for a relatively free word-order through obligatory case marking, although in practice a default word-order of SVO has emerged, with systematic deviations, triggered by complex (heavy) constituents and by pragmatics to express focus; these findings accord with relatively universal features found in natural languages (Jansen 2007). Cases are marked, viz. through a null-marking for the nominative and indirect object, an inflection in case of the accusative, and through a set of prepositions initially intended to be unambiguous (e.g., the English preposition "with" translates in two ways in Esperanto, through the instrumentalis "per" or with "kun," meaning together).

The qualification "relatively" is a commonly made one for the freeness of word-order. To be specific, it refers here to the fact that within constituents word-order is fixed for determiners, prepositions and negation & degree particles, while being free for adjectives and nouns. The order of constituents has a higher amount of freedom, but the order of prepositional phrases does reflect their place in argument structures. Lastly wh-question formation co-occurs with a word-order transformation, ie., if the wh-constituent is an object it is moved to sentence initial position. Strangely enough this does not happen for polar questions, which are marked with a polar question forming particle. This leaves Esperanto in the perhaps uncommon position of marking one type of question with word-order (which is desirable since the interrogative pronouns also serve

as relative pronouns), and the other with a particle.

We should also consider the ambiguities introduced by relaxing word-order. Since the accusative is marked through a declension with obligatory agreement, it is trivial to distinguish subjects and objects. However, boundaries between other constituents such as the nominative and the indirect object or the end of prepositional phrases are unmarked, and can result in ambiguity due to the aforementioned underspecification.

Concerning prepositions, the initial intention was to express some rather vague relations such as "believing *in* God" (which is neither spatial nor temporal, it would appear) with a semantically neutral preposition for an unspecified relation, the preposition "je"; however, this seems to have fallen in disuse, probably through interference from Ethnic languages. However, an interesting hypothesis could be that this reflects an evolutionary pressure for distinctions and ambiguities to correspond with the meanings that are actually expressed (the prior probability of wanting to express some meaning) – while an abstruse philosophical treatise may theoretically discuss "believing" while residing spatially or temporally "in God," this possibility is vanishingly rare so that making the distinction is wasted effort.

While Esperanto's morphology is agglutinative and synthetic[3], it is not poly-synthetic such as Inuit languages; single words cannot express what is denoted by a whole phrase in other languages, and grammatical roles are not marked, nor is the nature of the relation between elements that make up a word specified. It also does not feature incorporation such as in Catalan. Concerning the underspecification of relations between morphemes, consider the Dutch word "zoektechnieken", which could translated as "techniques for search," though "for" is not specified in the Dutch word. In an agglutinative language invariant morphemes that express only a single grammatical meaning are concatenated unmodified, such that identifying the elements that make up a word is relatively easy[4]. The process of word formation is completely productive and without exceptions; the only proviso is that a formation should make sense semantically when considering the meaning of its constituent elements (ie., the principle of compositionality modulo the Gricean maxim of manner).

There is obligatory agreement in number and declension within noun phrases. Verb paradigms are simple: tense is marked with the ending, person and number solely through the subject.

Esperanto's productive morphology can be summarized using a regular grammar. The following is adapted from Schubert (1993)[5], which in turn is based on Kalocsay's (1980) account. I have translated it into a regular grammar, proving that the word forms in the lexicon of Esperanto can be enumerated by a regular language; to my knowledge this is the first such description to date. The grammar for function words:

```
function_word := adverb | preposition | numeral
adverb := prefix adverb
preposition := prefix preposition
numeral := numeral numeral*
prefix := mal | ne | ...
```

[3]Esperanto has an index of agglutination of 1,0 and an average synthesis index (word-morpheme ratio) of 1.8-2, reported by Wells 1989

[4]ambiguities may arise through overlap; ie., when concatenating two smaller morphemes results in a string of characters that coincides with a larger morpheme

[5]caveat lector: Schubert incorrectly characterizes this grammar as recursive

```
suffix := il | et | ...
```

Content words are a little more involved (ibid):

```
 word := prefix* left* right ending (ε | declension)
left := right (ε | ending)
right := prefix* root suffix*
ending := o | a | e
declension := j | n
verb-ending := as | is | os | us | u
root := akv | far | ...
```

In these rules, "prefix" and "suffix" refers to a closed class of affixes; "(verb-)ending" refers to a one or two-character ending marking the Part-of-Speech; "declension" refers to either a null marking (nominative, singular) or the accusative and/or plurality marking. Furthermore, "*" is the Kleene star, "|" is the alternation operator, and lastly concatenation is implied. This grammar incorporates three processes of word-formation in Esperanto: derivation (concatenating elements to form words), compounding (concatenating elements to words to form more complex words), and POS category change. The latter refers to nominalizations and other possible mappings between Parts-of-Speech.

While this grammar should in all likelihood exhaust Esperanto's morphology, it is of little use for computational linguistics because of its ambiguity and flat structure. Whereas POS-tagging can be done practically error-free using a rule-based algorithm (save for proper names and foreign words), deeper morphological structure will depend on the morphemes in question, and possibly their semantics as well. However, in this project it is assumed that the latter does not play a major role as doing semantics is infeasible[6]. We will assume that derivations and compound words are constructed in a stochastic process that can be leared from examples (words with their appropriate structure, that is).

Another way in which the grammar falls short is that it does not consider the grammatical character of roots in Esperanto (Schubert 1993). Although initially controversial, the thesis that bare roots (without their grammatical endings) have a grammatical category to which they belong has by now been almost universally accepted in Esperantology. In effect this entails that roots in Esperanto belong to a prototypical semantic class (sometimes several). These classes are verbal, adjectival and noun-like; adverbs are part of the adjectival roots, arguably making up a "qualities" class. The typical example is "MARTEL" and "TOND", roots for hammer and cutting, respectively. The category of the former is a noun and thus "martelo" means a hammer, and the derived "marteli" means to hammer. The latter is a verbal root, with "tondi" meaning to cut, and the derivation "tondilo" meaning a tool to cut or a scissor, requiring an affix to denote a tool derived from a verb (directly affixing a noun ending to the root would mean "a cut"). Without recording the grammatical category of roots, a model of Esperanto morphology would not be able to predict the correct derivations and curtail overgeneration.

The present work glosses over a related feature of Esperanto roots, the fact that verbs are transitive or intransitive (valency), requiring an affix to change from the one to the other meaning. The reason for glossing over this aspect is that this information should become part of a more general account of argument structure (i.e., including prepositional arguments) that is beyond the scope of this project. Take these examples:

[6] it is my contention that semantics relies on extensive extra-linguistic world knowledge

(1) “La akvo bolas” (the water boils)
(2) “Mi boligas la akvon” (I boil the water)
(3) “Mi finis la libron” (I finished the book)
(4) “La libro finiĝis” (the book finished)

Sentences (1) and (3) contain the original verb, while (2) and (4) contain affixed verbs with a different subcategorization frame. This feature of Esperanto has been critized as being a needless distinction (common sense usually yields the correct meaning, as for example English demonstrates), as well as the rather arbitrary choices that have been made as to the transitivity, requiring a language user to memorize them by rote. It has also resulted in confusing paronyms such as “pesi” (to weigh something) and “pezi” (to weigh X kilos, to be heavy). It is however an unchangeable part of the language.

Previous work with Esperanto has resulted in a highly successful ($>$ 95% precision on a small test corpus) constraint grammar (Bick 2007), and a formal model of morphology and syntax in the form of an adpositional grammar (Gobbo 2009); an adpositional grammar is a dependency grammar combining directed dependencies with the dimension of trajector/landmark from construction grammar. These provide a means of comparison and a potential treebank.

Since there is no gold standard treebank with phrase structures for Esperanto, I will construct a small toy corpus for testing, as well as exploring a treebank generated with EspGram (Bick 2007) from a magazine corpus. Furthermore, experiments with U-DOP for both morphology and phrase structure are possible.

2.3 About Data-Oriented Parsing

Data-Oriented Parsing (Scha 1990; Bod & Scha 1996, henceforth DOP) is a computational framework for modeling natural language processing (NLP) and other hierarchical cognitive phenomena. Its basic assumptions are:

1. knowledge of language is made up of a corpus of concrete experiences rather than abstract rules; this concrete experience is stored in exemplars, pairings of surface forms and their structure.
2. when faced with a new sentence, all fragments of past experiences can be consulted to analyze the given sentence
3. fragments can be combined using one or more operations which obtain with a certain (estimated) probability

Two crucial aspects are the representation used to describe the concrete experiences and the method for ranking the possible analyses. Most research in Computational Linguistics currently focuses on isolated sentences annotated with phrase-structures trees; this project will follow the same approach with the addition of morphological structure. Various methods for selecting the best parse tree exist for DOP; the best performing methods combine a notion of simplicity (the derivation requiring the least amount of fragments) with likelihood (estimated probability); e.g., the most likely from the n shorted derivations.

It should be noted that Esperanto, as a free word-order language, is more suitably described using depedency structures. However, given extent of previous work on DOP with phrase-structure trees, I have opted to assume such hierarchical representantions instead. This is merely a pragmatically motivated assumption.

What makes DOP so promising is that if any computational approach to language can be said to successfully learn a language given enough data (ie., without recourse to innate knowledge), DOP is bound to be one of them. This is because the Data in Data-Oriented Parsing refers to exploiting all of the available data. Whereas more traditional methods in Computational Linguistics such as Probabilitistic Context-Free Grammars (PCFG) derive abstract rules from a treebank, throwing away valuable contextual information, DOP retains all exemplars and their fragments (modulo some potential pruning method corresponding to memory decay depending on usage and age etc.). This allows for the recognition of long-range dependencies such as in the construction "more X than Y." Also, compared to a PCFG, the statistical independence assumptions of DOP are weaker, because they can be spread over different derivations resulting in the same parse tree (ie., the assumptions made by each of the derivations of the most probable parse are corroborated by its other derivations). Prescher et al. (2004) observe that DOP combines the memory-based aspects of non-probabilistic machine learning techniques such as k-nearest neighbor with a probabilistic approach to deal with unseen (novel) exemplars; thus DOP provides a way to deal with the spectrum ranging from stock phrases that can be memorized by rote to completely novel sentences. The larger the fragments used in a derivation, the less independence assumptions need to be made; however, novel sentences can be parsed by backing off to smaller fragments. Thus, in the limit (as the corpus size approaches infinity) DOP does not make any independence assumptions at all.

2.4 Parallels to Data-Oriented Parsing

A fascinating parallel could be said to exist between DOP and the human immune system:

> "Edelman received the Nobel prize in 1972 for his model of the recognition processes of the immune system. Recognition of bacteria is based on competitive selection in a population of antibodies. This process has several intriguing properties (p. 78):
>
> 1. There is more than one way to recognize successfully any particular shape;
> 2. No two people have identical antibodies;
> 3. The system exhibits a form of memory at the cellular level (prior to antibody reproduction).
>
> Edelman extends this theory to a more general "science of recognition":
>
> By "recognition," I mean the continual adaptive matching or fitting of elements in one physical domain to novelty occurring in elements of another, more or less independent physical domain, a matching that occurs without prior instruction. [T]here is no explicit information transfer between the environment and organisms that causes the population to change and increase its fitness. (p. 74)" – Clancey (1991)

This general theory that is hinted at here is Edelman's Neural Darwinism, a theory of competition describing the development of the human

brain and the development of consciousness. The "species" selected for might be mental categories, conceptualizations, linguistic exemplars, etc.

DOP's notion of *spurious ambiguities* (different ways of deriving the same parse tree) accords perfectly with 1). While DOP does not explicitely claim that "no two people have identical [exemplars]", it might very well be (which dramatically changes the scope of DOP from a potentially purely linguistic account modeling a language to a necessarily psychological one modeling an idiolect); certainly no two individuals will have the exact same corpus. I am unsure exactly how to interpret 3), but reliance on memory is certainly the defining trait of DOP (as opposed to other formalisms which are typically biased to computation over memory).

2.5 DOP and Esperanto

I attribute the idea of applying Data-Oriented Parsing to Esperanto to Ken Miner (2006a):

> "Eĉ se ni disvolvus stokastajn alirojn bazitajn sur Datum-Orientita Pritraktado (DOP), ankorau necesus denaskaj parolantoj por validumi tiajn modelojn. Kiam temas pri la normala lingvistiko, ne eblas eskapi la neceson de denaskaj parolantoj kiel fina kontrolo."
>
> Even if we develop stochastic approaches based on Data-Oriented Parsing (DOP), native speakers would still be necessary for evaluating such models. When we speak of normal linguistics, it is impossible to escape the necessity of native speakers as the ultimate arbiters.

The quote is from a rather gloomy article on the lack of negative evidence for Esperanto, and the resulting impossibility of doing real linguistics (as opposed to the parochial "Esperantology"). Note that "native speakers" refers here specifically to speakers who use Esperanto in their day-to-day life with their peers, not in the broader sense of any language learned from parents. Native speakers in the latter sense exist but play a marginal role in the Esperanto movement, native speakers in the former sense do not exist and would violate the relative neutrality of Esperanto as an international language. I personally do not think this lack of evidence makes linguistics on Esperanto problematic, because gramaticality judgements and semantic intuitions are philosophically problematic no matter how many native speakers are available to supply them. While it is correct that there can be no negative evidence about the grammaticality or felicity of an Esperanto construction, the same goes for writing a poem in English: as long as the poem is required to be novel and original it needs to be composed with recourse to some creative "estimator" which judges whether a novel combination of words makes sense; positive evidence from corpora is of little value to this task, because it is biased to rehashing previously learned constructions, although it is undoubtedly a precondition as a language model. Such a creative estimator must prune the potentially infinite space of low probability events according to a subjective aesthetic ranking and threshold. Incidentally, it should be noted that Esperanto has a startingly rich tradition of translated and original poetry, ranging from its very inception up to the present day. Poetry has been one of the driving forces in coining neologisms, because of their affective connotations. It may be desired to express an antonymic meaning without evoking its opposite through the presence of its morpheme, compare

"malgaja" (un-cheerful, sad) and "trista"; similarly it may be dangerous to refer to "maldekstra" (left) in a noisy environment, as opposed to its proposed neologism "live."

The appropriateness of DOP for Esperanto should be noted. In contrast with the earlier a priori, philosophical languages published as completed projects (Maat 1999), Esperanto was presented in a modest brochure (Zamenhof 1887) purporting to fully describe its grammar in 16 rules, along with examples of original and translated prose and poetry, inviting the reader to start building and using the language by following its examples. That Zamenhof summarized his language in 16 rules may well have been a nod to the rival constructed language Volapük (Schleyer 1884), a popular but highly complex language of bygone days purportedly communicated to its author by God. The complexity of Volapük is demonstrated by the fact that its verb paradigm contains 1584 conjugations, by combining tense, aspect, voice, person, number and gender, among others. Such features made Volapük difficult to learn and use, just as the philosophical languages. During the first Esperanto congress the *Fundamento* (Zamenhof 1905) was ratified as the untouchable foundation of the language[7], containing the 16 grammar rules, a dictionary with 2600 words and translations in six languages, and a collection of exercises; all of these had been published at least a decade earlier and where already sanctioned through practice. In effect, the *Fundamento* can be considered as the authoritative corpus on Esperanto, to which only new vocabulary is to be added as needed, provided that it follows Esperanto orthography. Concerning morphology in particular, Schubert (1993) notes, after referring to Zamenhof instruction of consulting the supplied dictionary of roots and affixes:

> "Apart from this recipe for deciphering Esperanto texts, Zamenhof did not tell the users of his language exactly HOW to build complex words. He relied on providing a vast number of models and examples" (emphasis in the original)

Further on, Schubert notes:

> "Zamenhof may have intuitively felt the impossibility of describing a language exhaustively by means of rules. Such an insight would make his thinking very modern indeed. In any case he preferred to give examples rather than working out a detailed word grammar."

This clearly justifies our intention of analyzing Esperanto using an exemplar-based model, not only pragmatically because of DOP's success, but historically as well, since it accords with Esperanto's emergence. An interesting sidenote is that in the years after its publication, Esperanto's word formation processes appear to have regularized (Schubert 1989), favoring new coinings such as "aspekti" (to appear) over semantically opaque Germanisms such as "elrigardi" (Wennergren 2005) (literally to look out) in the sense of to appear (Dutch "er uitzien", German "aussehen"), naturally the literal sense of looking out e.g. a window remains. While such systematization may be reminiscent of creolization where a pidgin acquires a relatively complex rule system, it should be noted that Esperanto is neither a pidgin (since it has a grammar) nor a creole; an argument against the creolization of Esperanto is that creolization is by

[7] perusable online at http://www.akademio-de-esperanto.org/fundamento/index.html

definition driven by a newly formed, geographically homogeneous community of native speakers, which Esperanto certainly does not have. Furthermore, if Esperanto were to be a pidgin (it is not; cf. Haitao 2001), it would be one of an extremely curious sort: a pidgin with an authoritative corpus and a language academy overseeing its development. As Miner (2008) remarks, the latter is something which Chomsky could have facetiously remarked, but instead he has claimed (quite incorrectly) that Esperanto is not a language because it lacks a generative grammar[8], putatively because it "parasitizes" on other languages[9]; this clearly belies his ignorance of Esperanto (only its vocabulary is borrowed from European languages, its grammar is autonomous (Jansen 2007)), as well as being an obvious non-sequitur (perhaps Chomsky implicitly believes that *real* languages develop *de novo* without any interlinguistic interaction to speak of).

[8] A very debatable implicature is being made that natural languages do have a generative grammar, with all the assumptions that come with that

[9] paraphrased from an interview transcript available at http://www3.sympatico.ca/mlgr/chomsky.pdf

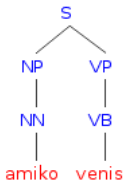


Figure 1: Translation: a friend came

3 Practice

3.1 Tag set

The tag set that I will use for the hand-annotated corpora, inspired by the Penn-treebank, is as follows:

- Constituents: VP, PP, NP, N' (constituents that behave like a noun), NC (conjunction + NN/N'), NPC (conjunction + NP), VPC (conjunction + VP), SC (conjuction + S), S' (if/that + S).
- Part-of-speech (simplified version of Penn tagset): NN, VB, PR, JJ, DT, RB, PRP, CC, UH
- Morphology, open class: N (noun), V (verb), J (adjectival), closed class: P (prefix), S (suffix), and auto-generated unique tags for all grammatical endings and declensions (o, j, n, etc.).

The Monato treebank uses a different tag set, based on the EspGram[10] constraint grammar. The POS tags of the morphology corpus should be adapted to fit those of the Monato treebank.

Figure 1 and 2 shows some annotated example sentences.

Some annotated example words are listed in figure 3.

3.2 Implementation

The implementation uses the Goodman (1996) reduction of DOP to a PCFG. I have written my own implementation[11], using NLTK (Bird et al., 2009). Future work should extend this implementation to add better estimators such as Backoff DOP or DOP*.

After producing a PCFG parsing is done using bitpar (Schmid 2004), an efficient bit vector based chart parser.

In order to apply the Goodman reduction to an arbitrary treebank, the reduction has been generalized to deal with arbitrary trees (not just trees in Chomsky normal form). This is done by translating subtrees of the form $(A\ B_1\ \ldots\ B_n)$ to rules of the form $A \rightarrow B_1 \ldots B_n$ with relative frequency:

$$\frac{\prod_{m=1}^{n} (freq(B_m) \text{ if } B_m \textit{ has an id else } 1)}{freq(A)}$$

[10] http://beta.visl.sdu.dk/visl/eo/index.php

[11] Full source code available at http://www.github.com/andreasvc/eodop, including the code dealing with morphology and preprocessing etc.

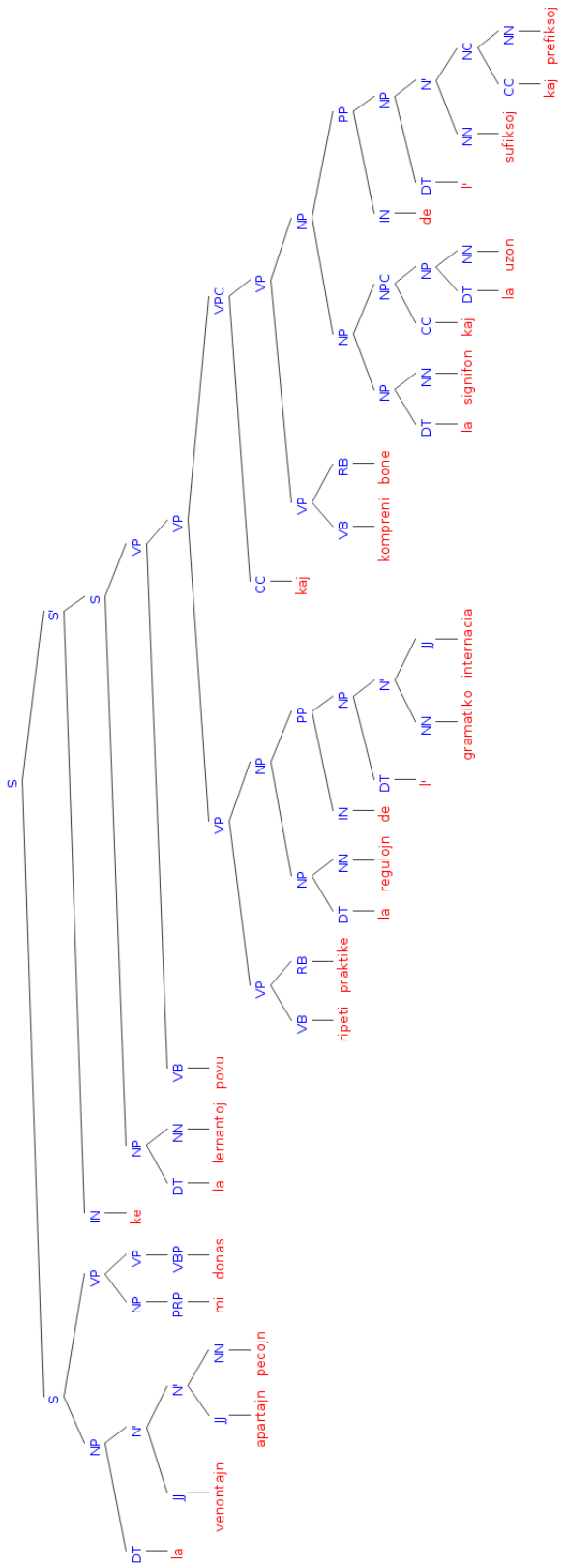


Figure 2: Translation: the following separate pieces I give so that the students will be able to rehearse practically the rules of the international grammar and understand properly the meaning and usage of the suffixes and affixes

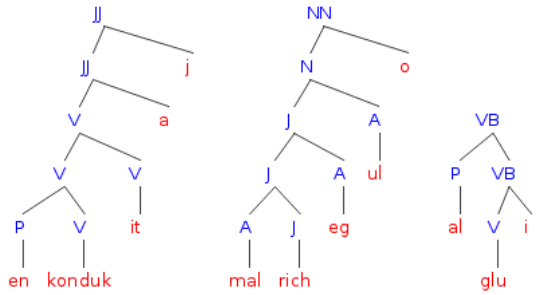


Figure 3: Sample words. Translations: those-that-were-introduced, very-poor-person, gluing-to

In order to fully separate terminals from non-terminals, all terminals are assigned an unique tag if they don't have one yet.

3.3 Segmentation

Before a morphological structure can be assigned to a word, it must be segmented into morphemes (similar to tokenization before parsing syntax). While it is claimed that in agglutinative languages in general and in Esperanto in particular it is "trivial" to recover the segments that make up a word (eg. Schubert 1993), this is a rather informal remark which is not borne out in practice. Morpheme boundaries are not marked, and ambiguities may arise due to overlapping roots. Hana (1998) notes a lexical homonymy rate of 13.6%.

I have devised a form of "Data-Oriented Segmentation" to expand the coverage of segmentation beyond that of the words in the morphology corpus. The algorithm works as follows:

1. take the set of segmented words in the corpus by reading off the leaves of their trees
2. construct a dictionary from positions to the set of morphemes occurring at that position
3. generate possible words by taking the cartesian product of all morphemes occurring at position 0 and 1, corresponding to all possible 2-morpheme words using the available vocabulary of roots.
4. repeat until position n where n is highest number of morphemes in the treebank to generate all possible words with $n + 1$ morphemes.

Unfortunately this algorithm suffers from overgeneration. This should be remedied by discarding any segmentations contradicting the initial set of (supervised) segmentations.

An alternative method of generating segmentations:

1. take the set of segmented words in the corpus by reading off the leaves of their trees, store words as tuples of morphemes.
2. construct a dictionary from number of morphemes to words with that number of morphemes
3. generate possible words with n morphemes for all suitable n by taking the pointwise cartesian product of all words with n morphemes; ie., `cartpi(zip(words[n]))`, which coresponds to:

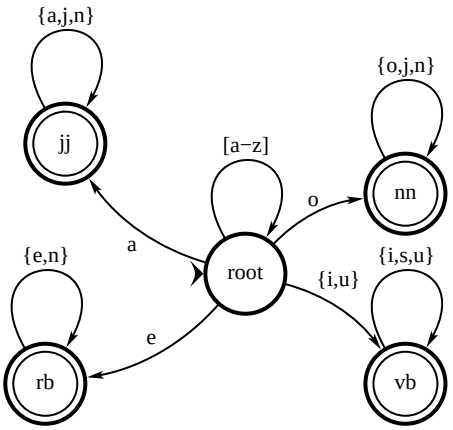


Figure 4: A sketch for a finite-state automaton for word classification in Esperanto.

$$\bigcup_{n=0}^{max(\{\,|x|\,:\,x\in words\})} \prod_{\{x\in words\,:\,|x|=n\}} x$$

This still overgenerates, though less so (eg., word class, plural and accusative endings in the wrong order; it may be necessary to treat endings separately). A third way would be to use a bigram model and produce every possible sequence up till a certain length, which avoids such issues.

When none of these approaches are able to segment a word we can resort to using the the context-free grammar described above, which recognizes any valid sequence of morphemes. The reason for using this as a last resort is that it does not distinguish between attested and unattested segmentations, let alone generalize over attested segmentations, as the previous methods do. The result will be that potential ambiguities that play no role in naturalistic data will rear their head.

In the current implementation I use neither the bigram nor the grammar for segmentation, as I have not yet encountered instances where it would be make a difference.

3.4 POS tags for unknown words

Since part-of-speech tags are transparently marked in Esperanto, it is possible to assign tags to any open class word. This can be done using a Finite-State Automaton such as in figure 4.

The actual automaton is more complicated, firstly because it is desirable to make it deterministic, secondly it can be made more strict because inflections may occur only once and in a fixed order.

A deterministic version replaces the ambigious transitions for the vowels with transitions back to the root state from the accepting states, and transitions between all the accepting states; for this reason such an automaton is given rather as a transition table in 1.

from	to	condition
root	rb	(e)
root	jj	(a)
root	nn	(o)
root	vb	(i,u)
root	root	([a-z])
vb	vb	(i,s,u)
rb	rb	(e,n)
jj	jj	(a,j,n)
nn	nn	(o,j,n)

Table 1: transition table for the deterministic finite-state word-class automaton. The start state is root, the final states are the open-class tags {nn, jj, rb, vb}. Strictly speaking special states should be created for accepting at most one declension ({j,n}) or conjugation ({i,u,s}), but that is overgeneration which is not as bad as undercoverage.

3.5 DOP model composition

In order to produce a combined morphology-syntax model, it is necessary to be able to compose a DOP model and a treebank. This is defined in the following manner:

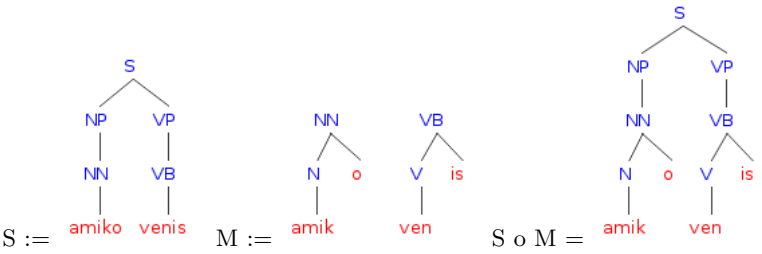
1. let M be a DOP model and S a treebank, where for example M contains morphology and S contains phrase structure trees.
2. the composition M o S yields a new DOP model by generating a new treebank S' based on the trees in the treebank S annotated with analyses of words parsed with M (assuming correct segmentation).
3. treebank S' is generated by iterating over the POS tags of the trees in S and substituting each POS tag with a tree from M.
4. the morphology-syntax model is obtained by instantiating a DOP model from S'.

Note that this procedure assumes that disambiguation of morphology is both context- and error-free; the most probable parse is used for decorating the syntax treebank[12]. This assumption should be empirically verified. Miner gives the example of a "fibestejo", which could be either a "fi(bestej)o" (a filthy place for animals), or a "fi(best)ejo" (a place for filthy animals). Another example he cites shows that some ambiguities can be semantically ruled out: an "eksklubano" is an ex-member of a club, not a member of an ex-club (because an ex-club cannot have members).

An example of the procedure:

```
S := { (S (NP (NN amiko)) (VP (VB venis))) }
M := { (NN (N amik) o), (VB (V ven) is) }
S o M = { (S (NP (NN (N amik) o)) (VP (VB (V ven) is))) }
```

[12]Because of technical limitations I have employed the n best parse trees to approximate the most probable parse



3.6 Corpora

Hand-annotated corpora:

- Morphology: hand annotated list of 290 words, containing all closed class words and affixes, and various open class roots and derivations. Compiled from various more or less naturalistic sources (e.g., Wennergren 2005, Miner 2006b).
- Syntax: hand annotated list of 14 sentences (first paragraph of Zamenhof's Dua Libro). Coverage of morphology is 100% with respect to this corpus. This should be extened to cover the whole Fundamento.

Treebanks:

- morphology: semi-supervised corpus generated from dictionaries (TBD)
- syntax: Monato treebank (Bick, personal communication, a corpus parsed with EspGram (Bick 2007). Number of sentences: 1995, tokens: 30,397, types: 9247. Average sentence length: 15.338. Resulting grammar is about 1 GB. Treebank requires preprocessing, a basic filter was applied to prune parse trees whose leaves do not agree with the original input sentence (be it because of a parse error in the original or an incorrect conversion); also, unique POS tags are inserted for punctuation; about 1500 trees remain afterwards.

4 Results

4.1 Results on toy corpora

Using a syntax and morphological corpus that do not contain the word "ven'as", but with a morphology model that can derive it from "don'as" and the past tense "ven'is":

```
sentence: amiko venas
morphology:
(NN (N amik) o) (p=0.00417101147028)
(VB (V ven) as) (p=0.000334168755221)
syntax:
error Grammar does not cover some of the input words: "'venas'".
morphology + syntax combined:
['amik', 'o', 'ven', 'as']
(S (NP (NN (N amik) o)) (VP (VB (V ven) as))) (p=1.12188584593e-28)
```

The corpus contains the plural "prefiksoj," which is inflected to an accusative here:

```
sentence: mi donas prefikson
morphology:
(PRP mi) (p=1.0)
(VB (V don) as) (p=0.0350877192982)
(NN (NN (N prefiks) o) n) (p=6.08906783983e-05)
syntax:
error Grammar does not cover some of the input words: "'prefikson'".
morphology + syntax combined:
['mi', 'don', 'as', 'prefiks', 'o', 'n']
(S
  (NP (PRP mi))
  (VP
    (VB (V don) as)
    (NP (NN (NN (N prefiks) o) n)))) (p=9.85999896556e-46)
```

However, it is perhaps unfair not to assign categories to unknown words. In the following results I let a deterministic finite state automaton assign the right POS tags to unknown words, and use a list of possible morpheme tags with uniform probabilities to tag unknown morphemes (for words with a single root the morpheme tagging will default to the POS tag marked by the ending, which will usually be correct).

Here is a large sentence from later in the "Dua Libro" (which is, fittingly, about word formation in Esperanto):

```
sentence: Vortoj kunmetitaj estas kreataj per simpla kunligado de simplaj vortoj
morphology:
Vortoj (NN (NN (N Vort) (NN_o o)) (NN_j j))
kunmetitaj (JJ (J (V kunmetit) (J_a a)) (JJ_j j))
estas (VB (V est) (VB_as as))
kreataj (JJ (J (V kreat) (J_a a)) (JJ_j j))
per (IN per)
simpla (JJ (J simpl) (JJ_a a))
kunligado (NN (J kunligad) (NN_o o))
de (IN de)
simplaj (JJ (J (V simpl) (J_a a)) (JJ_j j))
vortoj (NN (NN (N vort) (NN_o o)) (NN_j j))
morphology + syntax combined:
['Vort', 'o', 'j', 'kunmetit', 'a', 'j', 'est', 'as', 'kreat', 'a', 'j',
 'per', 'simpl', 'a', 'kunligad', 'o', 'de', 'simpl', 'a', 'j', 'vort', 'o', 'j']
(S
  (NP (NN (NN (N Vort) (NN_o o)) (NN_j j)))
  (VP
    (NP (JJ (J (V kunmetit) (J_a a)) (JJ_j j)))
    (VP
      (VB (V est) (VB_as as))
      (VP
        (JJ (J (V kreat) (J_a a)) (JJ_j j))
        (NP
          (NP
            (JJ (V (N per) (V simpl)) (JJ_a a))
            (NN (N kunligad) (NN_o o)))
          (PP
            (IN de)
            (N\'
              (JJ (J (V simpl) (J_a a)) (JJ_j j))
```

```
(NN (NN (N vort) (NN_o o)) (NN_j j)))))))))
```

See figure 5 for the tree.

There are some mistakes in segmenting (kun-met-it, kre-at, per simpl-a, kun-lig-ad). The phrase structure has mistakes as well, eg. "vortoj kunmetitaj" is a constituent, "per simpla..." should be a PP but this is overlooked because it got an incorrect POS tag. But given that the syntax corpus contains only 14 sentences it is perhaps striking that a parse was produced at all.

The modularist approach yields the following parse tree:

```
syntax \& morphology separate:
Vortoj kunmetitaj estas kreataj per simpla kunligado de simplaj vortoj
(S
  (NP
    (NN (NN (N Vort) (NN_o o)) (NN_j j))
    (JJ (J (V kunmetit) (J_a a)) (JJ_j j)))
  (VP
    (VP
      (VB (V est) (VB_as as))
      (NP (JJ (J (V kreat) (J_a a)) (JJ_j j)) (IN per)))
    (NP
      (NP (JJ (J simpl) (JJ_a a)) (NN (J kunligad) (NN_o o)))
      (PP
        (IN de)
        (N\'
          (JJ (J (V simpl) (J_a a)) (JJ_j j))
          (NN (NN (N vort) (NN_o o)) (NN_j j)))))))
```

The morphology is identical, but syntactically the results are a little different, eg. the first noun and adjective are together in an NP. However, the preposition "per" appears oddly at the end of an NP, instead of introducing a PP (in the previous tree it ended up prefixing an NP because the model cannot distinguish the difference between word and morpheme boundary).

That the finite state automaton is working can be seen from the following non-sense input:

```
sentence: tiadelaradon teluro didelas
morphology:
tiadelaradon (NN (NN (N tiadelarad) (NN_o o)) (NN_n n))
teluro (NN (N telur) (NN_o o))
didelas (VB (V didel) (VB_as as))
morphology + syntax combined:
['tiadelarad', 'o', 'n', 'telur', 'o', 'didel', 'as']
(S
  (NP (NN (NN (N tiadelarad) (NN_o o)) (NN_n n)))
  (VP (NP (NN (N telur) (NN_o o))) (VP (VB (V didel) (VB_as as)))))
syntax \& morphology separate:
(S
  (NP
    (NN (NN (N tiadelarad) (NN_o o)) (NN_n n))
    (NN (N telur) (NN_o o)))
  (VP (VB (V didel) (VB_as as))))
```

As can be seen, the words and roots receive the correct POS tags, which additionally is not derived from the default SVO order. The DOP

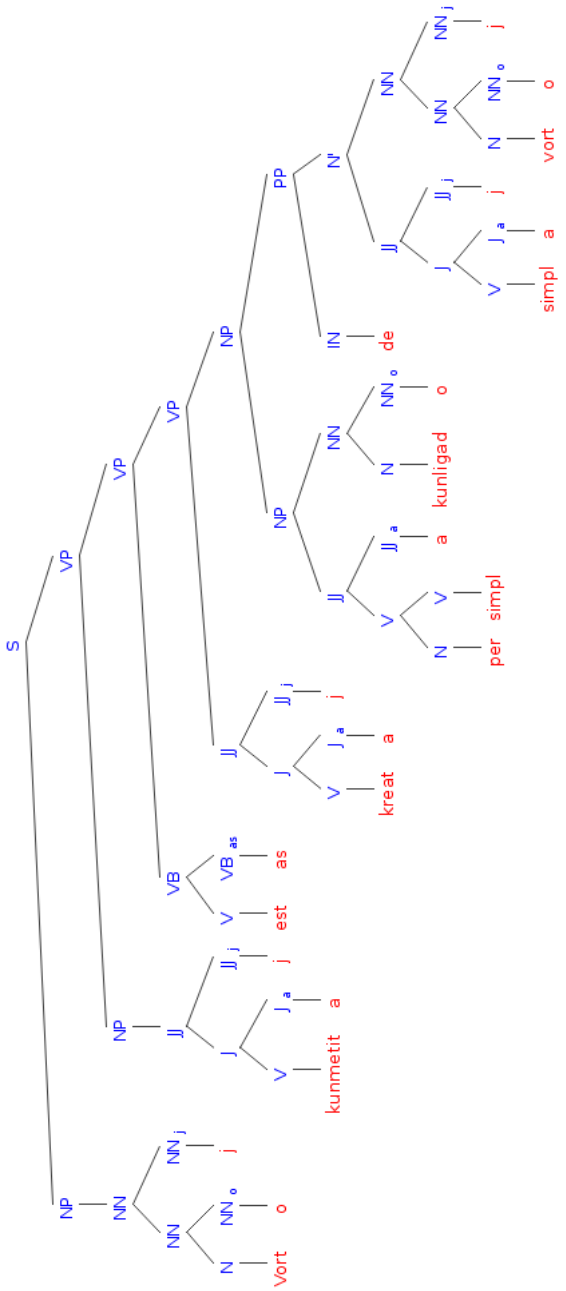


Figure 5: Derivation using the interactionist approach. Translation: Derived words are created using simple concatenation of simple words [NB: words means roots here]

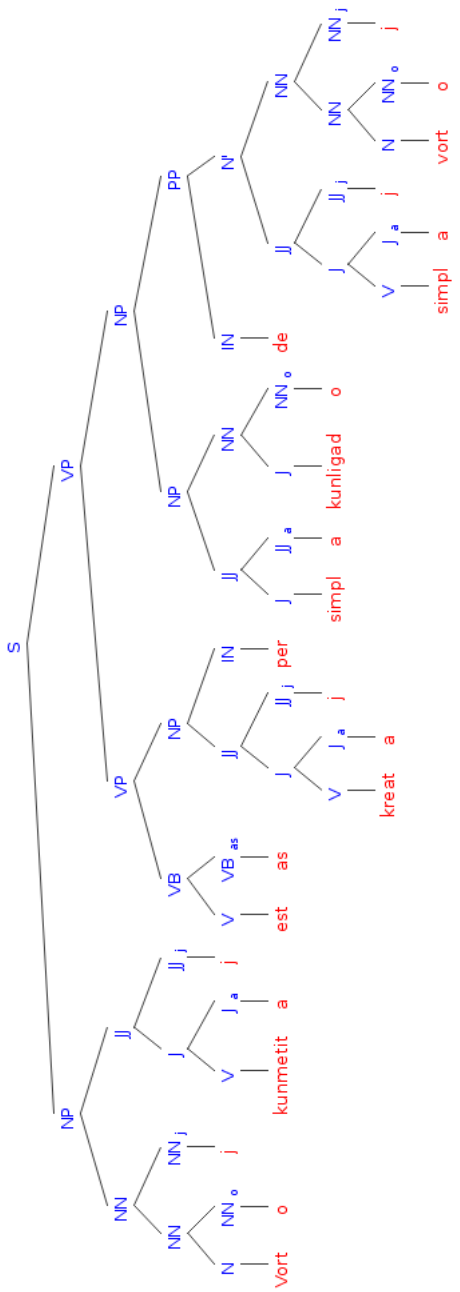


Figure 6: Derivation using the modularist approach, same sentence.

model where morphology is opaque to syntax considers the two nouns to be a single noun phrase, which could have been trivially excluded by attending to the morphology (but perhaps an accusative category label would have been enough).

4.2 Evaluation

Tenfold testing of Monato treebank, with and without regard for morphology.

To be done. Unfortunately parsing using the full monato corpus proved to be too much for bitpar, as it fails with a memory allocation error before producing any results. Alternatives would be to sample from the subtrees (ie., sample from the rules containing IDs plus all the rules without IDs), or using a different parser, one optimised more for memory usage.

4.3 Future work

Unknown words are already handled properly by the DFSA classifier supported by `bitpar` (save for proper names and foreign words, which are always problematic). However, dealing with unknown morphemes or morphemes used with an unattested tag is more difficult because morphemes are not explicitely marked, except by association with grammatical endings, which is not conclusive because the word may be the result of a category transformation (e.g., a nominalization). To deal with this problem I envision a two level Hidden Markov Model (HMM) for assigning tags to unknown words and morphemes. The first level will be words, where the hidden layer consists of their POS tags. The second layer will be the morphemes of those words, where the hidden layer consists of their morpheme tags. The HMM would be trained on the same training corpus as the DOP model, and its results should be added to the lexicon of the DOP model (ie., the rules of the form 'POS tag → terminal'). It is unclear to me how this affects the DOP probability model, but `bitpar` will certainly have no qualms with it as it expects frequencies anyway, so as to be able to do smoothing.

Another important improvement is to add explicit word boundaries when merging morphological structures with phrase structures. This can probably be implemented using a new root node containing the morphological analysis and a single space as children, but has not been explored yet.

Concerning morphology, it should be noted that e.g., Miner (2006b; and others cited therein) annotate all morphemes in complex words with their implied grammatical endings, such that a "vaporŝipo" (steamboat) is analyzed as "vaporo-ŝipo."[13] Such an analysis introduces additional ambiguity and departs from the empirical evidence, but it may be necessary for linguistic reasons. I have not considered it because I do not consider it realistic to assume such trace constituents, but there is no evidence against them other than their absence in most surface forms. On the other hand, the morphological structures that Miner presents are unlabeled, which makes it harder to generalize over multiple exemplars as categories are only represented in terminals. The choice to label morphological constituents with the same labels as individual roots seems useful and justifiable.

[13]Note that both surface forms are technically correct, but in practice the latter is only used for phonological reasons relating to prosody etc.

The Distributed Language Translation (DLT) project, which used Esperanto as a pivot language for Machine Translation, went as far as adding accusative endings inside complex words where implied. This goes against the grammar of Esperanto, but it is obviously useful for disambiguation.

5 Conclusion

We have described a regular grammar that enumerates the word forms of Esperanto's lexicon, which can be used to automatically segment word strings. Using a DOP model the resulting sequence of morphemes and tags can be analysed and assigned a hierarchical structure. The resulting DOP model can either be merged with a syntactic treebank into a combined DOP model, or mapped to the leaves of the parse trees produced by a syntactic model, to obtain tree structures with both phrasal and morphological constituents.

We described an implementation using NLTK of the Goodman reduction that is generalized to arbitrary trees, which outputs a grammar that can be parsed by the efficient chart parser Bitpar. Using a list of open class tags and a finite state automaton we can assign tags to unknown words and morphemes.

The resulting system has been applied to a small corpus of morphology and syntax, hinting at the advantage of merging morphology and syntax treebanks before constructing a DOP model. Evaluation with a larger syntactic treebank, as well as the induction of morphology tags from dictionaries remains to be done. However, the groundwork for such an enterprise has been laid, as well as for addressing the research questions. It seems likely that morphology will be crucial for parsing syntax, especially when morphology is rich and highly structured.

∞

6 References

Bick, Eckhard (2007), "Tagging and Parsing an Artificial Language: an annotated web-corpus of Esperanto," in: *Proceedings of Corpus Linguistics*, Birmingham, UK.
http://beta.visl.sdu.dk/pdf/CorpusLinguistics2007_esp.pdf

Bird, Steven, Edward Loper & Ewan Klein (2009). "Natural Language Processing with Python." O'Reilly Media Inc.

Bod, Rens & Scha, Remko (1996) "Data-Oriented Language Processing: an overview." Research reports, Institute for Logic, Language and Computation, University of Amsterdam.
http://dare.uva.nl/document/1144

Clancey, W.J. (1991), "The biology of consciousness: Comparative review of Israel Rosenfield, The Strange, Familiar, and Forgotten: An anatomy of Consciousness and Gerald M. Edelman, Bright Air, Brilliant Fire: On the Matter of the Mind," *Artificial Intelligence* vol. 60, pp. 313–356

Gobbo, Federico (2009), "Adpositional Grammars: a multilingual grammar formalism for NLP," PhD dissertation, Universita degli Studi dell'Insubria.

Goodman, Joshua (1996), "Efficient Algorithms for Parsing the DOP Model". *Proceedings Empirical Methods in Natural Language Processing*, pp. 143-152.
http://acl.ldc.upenn.edu/W/W96/W96-0214.pdf

Jackendoff, Ray (2003), "Prcis of Foundations of Language: Brain, Meaning, Grammar, Evolution," Behavioral and Brain Sciences (2003), 26:6:651-665 Cambridge University Press.

Jansen, W. (2007). "Woordvolgorde in het Esperanto: normen, taalgebruik en universalia" (Word-order in Esperanto: norms, usage and universals). PhD thesis, LOT Utrecht.

Jurafsky, D. & Martin, J.H. (2000), "Speech & Language Processing An introduction to natural language processing, computational linguistics, and speech recognition," Pearson Education.

Hana, Jiři, "Two-level morphology of Esperanto," MSc thesis, Charles University Prague, Faculty of Mathematics and Physics. http://www.ling.ohio-state.edu/~hana/esr/thesis.html

Haitao, Liu (2001), "Creoles, Pidgins, and Planned Languages." Interface. Journal of Applied Linguistics / Tijdschrift voor Toegepaste Lingustiek 15 [2]. pp. 121–177.

Kalocsay, Kálmán & Waringhien, Gaston (1980), Plena Analiza Gramatiko de Esperanto (Complete, analyzed Grammar of Esperanto), Rotterdam, Universala Esperanto-Asocio.

Maat, Jaap (1999), "Philosophical Languages in the Seventeenth Century: Dalgarno, Wilkins, Leibniz," Amsterdam, Institute for Logic, Language and Computation.

MacWhinney, B. (1987), "Mechanisms of Language Acquisition," Lawrence Erlbaum Associates, NJ.

Miner, Ken (2006a), "Tranchitaj frazoj kaj la probleme pri negative evidento" (Cut phrases and the problem of negative evidence). March 2006.
http://www.sunflower.com/~miner/NEGATIVA_package/negativa.html

Miner, Ken (2006b), "Rimarkoj pri 'En la komenco estas la vorto' de Geraldo Mattos (fina versio)," (Comments on 'In the beginning was the word' by Geraldo Mattos (final version)).
http://www.sunflower.com/~miner/EKVO_package/ekvo.html

Miner, Ken (2008), "La neebleco de priesperanto lingvoscienco," (The impossibility of Esperanto linguistics). October 2008.
http://www.sunflower.com/~miner/LINGVISTIKO_package/lingvistiko.html
Also published in "La arto labori kune: festlibro por Humphrey Tonkin" (The art of working together: Festschrift for Humphrey Tonkin). Roterdam, Universala Esperanto Asocio, January 2010

Pinker, S. (1994). The language instinct: How the mind creates language. New York: W. Morrow.

Prescher, D., Scha, R., Sima'an, K., Zollmann, A., (2004) "On the statistical consistency of DOP estimators." In *Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands*, Antwerp, Belgium.

Scha, Remko (1990), "Taaltheorie en Taaltechnologie; Competence en Performance" (Language theory and language technology: Competence and Performance), in Q.A.M. de kort and G.L.J. Leerdam (eds.), *Computertoepassingen in de Neerlandistiek* pp. 7-22, Almere: Landelijke Vereniging van Neerlandici (LVVN-jaarboek). English translation http://www.hum.uva.nl/computerlinguistiek/scha/IAAA/rs/cv.html

Schleyer, Johan Martin (1884), "Volapük. Grammatik der Universalsprache für alle gebildete Erdbewohner," Überlingen am Bodensee: Buchdruckerei August Feyel, Buchhandlung Aug. Schoy. Third edition.

Schmid, Helmut (2004), "Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors," *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004), Geneva, Switzerland. http://www.ims.uni-stuttgart.de/www/projekte/gramotron/PAPERS/COLING04/BitPar.pdf

Schmid, Helmut, Arne Fitschen and Ulrich Heidi (2004), SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection, Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004), p. 1263-1266, Lisbon, Portugal. http://www.ims.uni-stuttgart.de/www/projekte/gramotron/PAPERS/LREC04/smor.pdf

Schubert, Klaus, 1989. "An unplanned development in planned languages", en Klaus Schubert, red., Interlinguistics: Aspects of the Science of Planned Languages [= Trends in Linguistics: Studies and Monographs 42], Mouton de Gruyter.

Schubert, Klaus (1993), "Semantic compositionality: Esperanto word-formation for language technology." *Linguistics* 31: 311-365.

Tsarfaty, Reut (2010). "Relational-Realizational Parsing," PhD thesis, Institute for Language, Logic and Computation (ILLC), University of Amsterdam.

Wells, John (1989), "Lingvistikaj aspektoj de Esperanto," Universala Esperanto Asocio, Rotterdam. Second edition.

Wennergren, Bertilo (2005), "Plena Manlibro de Esperanta Gramatiko," (Complete handbook of Esperanto Grammar), version 13.0, 14th of April 2005. Available online at http://bertilow.com/pmeg/.

Zamenhof, Dr. L. L. (1887/1968), "Internationale Sprache. Vorrede und Vollständiges Lehrbuch," Warschau, photographic reprint from 1968 (Saarbrücken: Artur E. Illtis). German translation of the original Russian brochure.

Zamenhof, Dr. L. L. (1905/1963), "Fundamento de Esperanto." Ninth edition with Introduction, Notes and Linguistics comments, edited by Dr. A. Albault (Esperantaj Francaj Eldonoj: Marmande, 1963).

Zollmann, Andreas & Sima'an, Khalil (2005), "A Consistent and Efficient Estimator for DOP." *Journal of Automota Languages and Combinatorics* vol. 10, pp. 367. http://staff.science.uva.nl/~simaan/D-Papers/JALCsubmit.pdf