Extracting and Annotating Wikipedia Sub-Domains*

— Towards a New eScience Community Resource —

Gisle Ytrestøl♣, Dan Flickinger♠, and Stephan Oepen♣♠

♣ University of Oslo, Department of Informatics

♠ Stanford University, Center for the Study of Language and Information

gisley@ifi.uio.no, danf@stanford.edu, oe@ifi.uio.no

Abstract

We suggest a simple procedure for the extraction of Wikipedia sub-domains, propose a plain-text (human and machine readable) corpus exchange format, reflect on the interactions of Wikipedia markup and linguistic analysis, and report initial experimental results in parsing and treebanking a domain-specific sub-set of Wikipedia content.

1 Motivation and a Long-Term Vision

Linguistically annotated corpora—for English specifically the Penn Treebank (PTB; Marcus, Santorini, & Marcinkiewicz, 1993) and derivatives—have greatly advanced research on syntactic and semantic analysis. However, it has been observed repeatedly that (statistical) parsers trained on the PTB can drop sharply in terms of parsing accuracy when applied to other data sets (Gildea, 2001, inter alios). Ever since its release, there have been concerns about the somewhat idiosyncratic nature of the PTB corpus (primarily Wall Street Journal articles from the late 1980s), in terms of its subject matter, genre, and (by now) age. Furthermore—seeing the cost of initial construction for the PTB, and its still dominant role in data-driven natural language processing (NLP) for English—design decisions made two decades ago perpetuate (sometimes in undesirable ways) into contemporary annotation work. PropBank (Palmer, Gildea, & Kingsbury, 2005), for example, performs semantic annotation on the basis of PTB syntactic structures, such that a discontinuous

*This report on work in progress owes a lot to prior investigation by Woodley Packard, who started parsing Wikipedia using the ERG as early as 2003. We are furthermore indebted to Francis Bond, Yusuke Miyao, and Jan Tore Lønning, for their encouragement and productive comments. The WeScience initiative is funded by the University of Oslo, as part of its research partnership with Stanford's Center for the Study of Language and Information.

structure like *Gulf received a takeover bid from Simmons of $50 million.* (simplified from WSJ0178) leads to an analysis of *receive* as a four-place relation (with a dubious ARG4-of role). Quite generally speaking, richly annotated treebanks that exemplify a variety of domains and genres (and of course languages other than English) are not yet available. And neither are broadly accepted gold-standard representations that adequately support a range of distinct NLP tasks and techniques.

In response to a growing interest in so-called eScience applications of NLP—computationally intensive, large-scale text processing to advance research and education—a lot of current research targets scholarly literature, often in molecular biology or chemistry (Tateisi, Yakushiji, Ohta, & Tsujii, 2005; Rupp, Copestake, Teufel, & Waldron, 2007; inter alios). Due to the specialized nature of these domains, however, many NLP research teams—without in-depth knowledge of the subject area—report difficulties in actually 'making sense' of their data. To make eScience more practical (and affordable for smaller teams), we propose a simple technique of compiling and annotating domain-specific corpora of scholarly literature, initially drawing predominantly on encyclopedic texts from the community resource Wikipedia.[1] Adapting the Redwoods grammar-based annotation approach (Oepen, Flickinger, Toutanova, & Manning, 2004) to this task, we expect to construct and distribute a new treebank of texts in our own field, Computational Linguistics, annotated with both syntactic and (propositional) semantic information—dubbed the WeScience Treebank. Should this approach prove feasible and sufficiently cost-effective, we expect that it can be adapted to additional document collections and genres, ideally giving rise—over time—to an increased repository of 'community treebanks', as well as greater flexibility in terms of gold-standard representations.

In an initial experiment, we gauge the feasibility of our approach and briefly discuss the interaction of (display) markup and linguistic analysis (§ 2 and § 3); we further report on a very preliminary experiment in sentence segmentation, parsing, and annotation (§ 4), and conclude by projecting these into an expected release date for (a first version) of WeScience.

2 Wikipedia—Some Facts and Figures

Wikipedia represents probably the largest and most easily accessible body of text on the Internet. Under the terms of the open-source GNU Free Documentation License, Wikipedia content can be accessed, used, and redistributed freely. Wikipedia

[1] In 2007 and 2008, the interest in Wikipedia content for NLP research has seen a lively increase; see `http://www.mkbergman.com/?p=417` for an overview of recent Wikipedia-based R&D, most from a Semantic Web point of view.

to date contains more than 1.74 billion words in 9.25 million articles, in approximately 250 languages. English represents by far the largest language resource, with more than 1 billion words distributed among 2,543,723 articles.[2] Its size, hypertext nature, and availability make Wikipedia an attractive target for NLP research.

Wikipedia's editing process distinguishes it from most other documents that are available on-line. An article may have countless authors and editors. In April 2008, the English Wikipedia received 220,949 edits a day, with a total of 175,884 distinct editors that month. Wikipedia text provides relatively coherent, relatively high-quality language, but it inevitably also presents a comparatively high degree of linguistic (including stylistic) variation. It is thus indicative of dynamic, community-created content (see below).

2.1 Domain-Specific Selection of Text

Our goal in the WeScience Treebank is to extract a sub-domain corpus, targeting our own research field—NLP. To approximate the notion of a specific sub-domain in Wikipedia (or potentially other hyper-linked electronic text), we start from the Wikipedia category system—an optional facility to associate articles with one or more labels drawn from a hierarchy of (user-supplied) categories. The category system, however, is immature and appears far less carefully maintained than the articles proper. Hence, by itself, it would yield a relatively poor demarcation of a specific subject area.

For our purposes, we chose the category *Computational Linguistics* and all its sub-categories—which include, among others, *Natural Language Processing*, *Data Mining* and *Machine Translation*—to activate an initial seed of potentially relevant articles. Altogether, 355 articles are categorized under *Computational Linguistics* or any of its sub-categories. However, some of these articles seemed somewhat out-of-domain (see below for examples), and several are so-called stub articles or very specific and short, e.g. articles about individual software tools or companies. It was also apparent that many relevant articles are not (yet) associated with either of these categories. To compensate for the limitations in the Wikipedia category system, we applied a simple link analysis and counted the number of cross-references to other Wikipedia articles from our initial seed set. By filtering out articles with a comparatively low number of cross-references, we aim to quantify the significance (of all candidate articles) to our domain, expecting to improve both the recall and precision of sub-domain extraction.

[2] Given the highly dynamic nature of Wikipedia, these statistics evolve constantly. We report on the stable 'release' snapshot dated July 2008, which also provides the starting point for our sub-domain extraction. See `http://en.wikipedia.org/wiki/Wikipedia` for up-to-date statistics on Wikipedia.

Among the articles that were filtered out from our original set of seed articles, we find examples like *AOLbyPhone* (0 references) and *Computational Humor* (1 reference). New articles, differently categorized, were activated based on this approach. These include quite prominent examples like *Machine learning* (34 references), *Artificial intelligence* (33 references) and *Linguistics* (24 references). Of the 355 seed articles, only 30 articles remain in the final selection. Confirming our expectations, filtering based on link analysis eliminated the majority of very narrowly construed articles, e.g. specific tools and enterprises.

However, our link analysis and cross-reference metric also activates a few dubious articles (in terms of the target sub-domain), for example *United States* (8 references). We have deliberately set up our sub-domain extraction approach as a fully automated procedure so far, avoiding any elements of subjective judgment. However, we expect to further refine the results based on feedback from the scientific community.

To suppress the linguistically less rewarding stub articles, we further applied a minimum length threshold (of 2,000 characters, including markup) and—using a minimum of seven incoming cross-references—were left with 100 Wikipedia articles and approximately 270,000 tokens.

2.2 Spelling Conventions and Stylistic Variation

English Wikipedia does not conform to one specific (national) spelling convention, but according to the guidelines for contributors and editors, the language within an article should be consistent with respect to spelling and grammar (thus, *centre* and *center* should not be used side-by-side). Articles are not tagged according to which variant of English they use, and we have yet to gather statistics on the distribution of English variants.

In our view, the WeScience Treebank reflects the kind of content that is growing rapidly on the Internet, namely community-created content. In such content one will typically expect more variation in style (and quality), more spelling mistakes and 'imperfect' grammar, as well as some proportion of text produced by non-native speakers. The WeScience Treebank may provide a useful point of reference for the study of the distribution of such phenomena in community-created content.

3 Wikipedia Markup

Wikipedia articles are edited in the *Wiki Markup Syntax*, a straightforward logical markup language that facilitates on-line rendering (as HMTL) for display in a web browser. Again, there are Wikipedia guidelines and conventions for how to edit or

add content, aiming to keep the architecture and design as consistent as possible. In preparing the WeScience Treebank we aim to strike a practical balance between (a) preserving all linguistic content, including potentially relevant markup, and (b) presenting the corpus in a form that is easily accessible to both humans and NLP tools. Thus, we define a textual, line-oriented WeScience exchange format (see Section 3.2 for a brief discussion of related, XML-based efforts and our rationale for choosing a plain-text format). From the raw source files of Wikipedia articles, we eliminate markup which is linguistically irrelevant—some meta information or in-text image data, for example—but aim to preserve all markup that may eventually be important for linguistic analysis. Markup indicating bulleted lists, (sub)headings, hyper-links, or specific font properties, for example, may signal specialized syntax or use – mention contrasts.

Following are some examples of Wikipedia markup that we want to preserve, because it closely interacts with 'core' linguistic content:

(1) [10120240] |* Design of [[parser]]s or [[phrase chunking|chunkers]]
for [[natural language]]s

(2) [10621290] |For example, in the following example, ″one″
can stand in for ″new car″.

The WeScience Treebank provides gold-standard 'sentence' boundaries (sometimes sentential units are not sentences in the linguistic sense; see below) with unique sentence identifiers. Examples (1) and (2) show the actual WeScience file format, where each sentence is prefixed by its identifier and the '|' separator symbol. In (1), the initial '*' indicates items in a bulleted list (which can exhibit various specialized syntactic patterns) and the square brackets show Wikipedia hyperlinks. Example (2), on the other hand, shows the use of *italics* (the interpretation of the double apostrophe in Wikipedia markup) for the purpose of quoting; i.e. the use – mention distinction is made as a font property only. We further discuss the interactions of display markup and syntactic structures in Section 4.2 below.

3.1 From Source Markup to the WeScience Corpus

Unwanted markup and entire sections from the source articles have been automatically removed, mainly by the use of a large number of regular expressions. These are parts of the articles which, as we see it, are irrelevant to linguistic analysis, including entire sections—like *See Also*, *References*, or *Bibliography*—or links to images, comments made by other users, and various Wikipedia-internal elements.

Once reduced to what we consider (potentially) relevant linguistic content, we applied semi-automated sentence segmentation. In a first, automated step, all linebreaks were removed from the original source text, and the open-source package

`tokenizer`[3] was used to insert sentence boundaries. This is a rule-based tool which proved very capable as a sentence segmenter. The procedure was further optimized by some customization, based on a manual error analysis. Most of the errors made by the segmenter could be attributed to 'misleading' (remaining) markup in its input (`tokenizer`, by default, expects 'pure' text). For instance, the tool initially failed to insert segment boundaries between some numbered list elements (where the Wikipedia markup '#', in a sense, takes on the function of sentence-initial punctuation). We have augmented our pre-processing pipeline with regular expressions that force segment boundaries in cases where `tokenizer` failed. Furthermore, we are currently experimenting with an additional layer of 'temporary simplification' for the sentence segmentation step. For markup elements that can be asserted to never span segment boundaries (Wikipedia links, `<source>` and `<code>` blocks, for example), segments of original text are temporarily replaced with simplified, placeholder tokens. Once sentence segmentation is complete, these replacements are reverted.

A second round of error analysis on a sub-set of 1,000 segments suggests a residual error rate of about 4 per cent, with half of these text preparation errors due to incomplete handling of wiki mark-up (e.g. for `<math>` and `<code>` blocks, colons marking indentation, and some hyperlinks). The other half of these errors are due to missing or spurious sentence breaks (often due to unusual punctuation clusters), and to confusion of picture captions or section headers with main text. After one more round of tuning of these preparation scripts, we will manually inspect and correct segment boundaries as we treebank.

The WeScience exchange format presents one sentence per line, in (mostly) plain text form. Unique eight-digit sentence identifiers provide ease of reference and are coded in a form that directly points back to the original source article (first three digits). The approximately 270,000 tokens in the corpus in our current selection amount to about 14,000 sentences, ranging up to 185 tokens in length (excluding markup), with a relatively dense distribution up to about 50 tokens (e.g. 107 sentences are between 50 and 55 tokens long). Sentences are consecutively distributed across 16 files ('sections' of a sort), where each section comprises up to 1,000 sentences, and no article is split between two files.

3.2 Related Initiatives

There are of course numerous other NLP initiatives who look at Wikipedia text as a corpus. One such resource is the WikiXML initiative at the Information and

[3] See `http://www.cis.uni-muenchen.de/~wastl/misc/` for background and access.

Language Processing Systems group of the University of Amsterdam.The project web site summarizes:

> Our XML version of Wikipedia was designed to serve as a multi-lingual text collection for experiments in Information Retrieval and Natural Language Processing, in particular, in the context of Cross-Language Evaluation Forum (CLEF). (`http://ilps.science.uva.nl/WikiXML/`)

Besides providing XML snapshots of Wikipedia in various languages, WikiXML offers a conversion tool that can be used to convert articles from Wikipedia Markup to XML format. In a sense, this step can be viewed as making article and markup more explicit, forcing it into the more rigid scheme of a validated XML document. The Amsterdam initiative is just one of several projects which convert Wikipedia native format to XML, typically with the explicit or implied goal of easing NLP tasks on Wikipedia text. WikipediaXML (Denoyer & Gallinari, 2006), for example, is another closely related initiative.

Although projects like these clearly pull in the same direction as ours with respect to preparing Wikipedia-based NLP corpora, we have deliberately decided against XML markup and, thus, were left developing our own preprocessing and conversion tools. First, in principle Wikipedia markup is no less explicit than XML, nor are document validation techniques restricted to XML (the Wikipedia server infrastructure, in a sense, validates Wikipedia markup every time it renders an article for on-line browsing). But largely owed to its compactness and design aiming to facilitate source-level editing, we find that Wikipedia markup strikes a better balance between machine and human readability.

Second, and more importantly, the central unit of analysis in our setup is the sentence. For increased flexibility in manipulating WeScience files, we want it to be the case that both individual sentences and any concatenation of sentences are valid structural units. Thus, in the tradition of the Un*x operating system, we have opted for a textual, line-oriented exchange format. In XML, on the other hand, there is no straightforward way of concatenating multiple documents into a new, valid document. In treebanking Wikipedia text, the document-level token structure and linguistic tokenization need not always be compatible. In example (1) above, for example, the sub-string *[[natural language]]s* would likely be considered two document tokens (one being a Wikipedia link, with a juxtaposed string token *s*), but in terms of syntactic structure *languages* will have to be a single word form. Where Wikipedia markup can leave document token structure implicit, our parsing and treebanking machinery (see Section 4) keeps track of linguistic token positions in terms of character position stand-off pointers. Hence, it is possible for *languages* to span a character range that does not coincide with document-level token boundaries.

4 Initial Parsing and Treebanking Results

In annotating the WeScience Treebank, we plan to apply the grammar- and discriminant-based approach of Oepen et al. (2004)—the open-source LinGO Redwoods environment, based on the LinGO English Resource Grammar (ERG; Flickinger, 2000).[4] In the Redwoods approach, the treebank records the complete syntacto-semantic analyses provided by the grammar (in the HPSG framework), coupled with tools to extract different kinds of linguistic representation (at variable granularity)—primarily labeled syntax trees, 'deep' dependency structures, and logical-form meaning representations. Annotation in Redwoods amounts to disambiguation among the candidate analyses proposed by the grammar and, of course, analytical inspection of the final result. To gauge the feasibility (and scalability) of this approach, we performed an initial 'blind' experiment, applying the ERG to the complete WeScience corpus and asking an expert linguist to disambiguate two of the 16 sections. To the best of our knowledge, the ERG has not been previously adapted for Wikipedia text, hence it is to be expected that—ideally in joint work with the LinGO team—grammar and parser performance can be further improved.

4.1 An 'Out-of-the-Box' Experiment

For this initial experiment, we constructed another layer of regular expressions (thirteen in total), to simply eliminate Wikipedia markup prior to parsing.[5] Basic parsing coverage of the corpus with the ERG reached 86 percent, with variation of up to 5 per cent in either direction for any one of the 16 sections. For each sentence, we recorded up to 500 highest-ranked analyses, using a pre-existing parse-selection model (for a different domain, viz. hiking instructions). In this initial experiment, we also imposed relatively restrictive resource limits on the time and memory usage permitted for any one sentence (a maximum of one gigabyte in process size or one minute in cpu time). The average number of tokens in each sentence is 17.9, and (using this specific configuration) we observed average parse times per sentence (to produce up to 500 analyses for treebanking) of just below five seconds.

We were positively impressed with comparatively good parsing coverage in this 'out-of-the-box' experiment, applying the ERG to a new genre and subject area.[6]

[4] The Redwoods approach is essentially a scaled-up adaptation of the techniques originally proposed by Carter (1997). Grammar- and discriminant-based annotation has been applied successfully to multiple languages and frameworks, for example by Bouma, van Noord, & Malouf (2001) and Rosén, De Smedt, & Meurer (2006), inter alios.

[5] For reasons discussed briefly in § 3 above, this simple-minded approach may inhibit successful analysis in some cases. In future experiments we aim to collaborate with the ERG developers on a genuine integration of pre-processing and parsing (using the framework of Adolphs et al., 2008).

[6] The grammar includes an unknown word facility, postulating underspecified lexical entries for

From earlier Redwoods reports, however, it is to be expected that basic parse success does not necessarily indicate the existence of a *correct* analysis. To determine the expected proportion of invalid analyses, we manually treebanked (using the Redwoods discriminant machinery) all inputs that parsed from two sections of the 16 sections. This exercise indicated a (relatively high) rejection rate of nearly 30 percent, i.e. during treebanking the annotator only found a correct analysis for a little more than 70 percent of all sentences.[7]

Combining 14 percent parse failures and a 30-percent rejection rate during treebanking, our prelimary experiment arrives at a fully correct analysis for about 60 percent of the segments in the WeScience corpus. These analyses are 'correct' according to the linguistic assumptions of the HPSG framework and its implementation in the ERG, and as such they provide comparatively fine-grained syntacto-semantic information—at moderate annotation cost. To further put these results into perspective, and of course to estimate to what degree one can hope to reduce treebank 'gaps' (introduced by either parse or annotation failure), we applied a first manual error analysis to the same two treebanked sections. This process revealed a number of distinct sources of failure, grouped as either shortcomings in *corpus preparation* or *analysis errors*. Table 1 shows a more detailed break-down of failure types, including a rough estimate (based on our selection of two sections) of the percentage of items affected in the full corpus.

4.2 Integrating Markup and Syntactic Analysis

As pointed out earlier, in some cases markup properties directly affect linguistic analysis. A prominent such example is the use of italics in a function similar to quotation (drawing use – mention distinctions or demarcating foreign language material), as we saw in example (2) above. To abstract from the concrete syntax of a specific markup language (like Wikipedia, HTML, or LaTeX), we have started to augment the ERG analysis grammar with selected elements of an *abstract* Grammatical Markup Language (dubbed GML). During input pre-processing for the parser, for example, italicized words or phrases are enclosed in 'opening' and 'closing italics' tokens: ¦*i new car i*¦, for part of example (2).

Parser-internally, GML tokens like these are treated much like punctuation marks, i.e. in the approach of Adolphs et al. (2008) such tokens are re-combined

open class categories on the basis of a standard PoS tagger. On the other hand, the configuration we used does *not* include additional robustness measures, i.e. the parser will fail in case there is no complete analysis, spanning the full input string.

[7] For comparison, Oepen et al. (2004) report a ten percent rejection rate. However, in their exercise they were treebanking considerably easier text (transcribed task-oriented dialogues), and working in a domain for which the ERG had been carefully adapted already.

Corpus Preparation	
1%	missing sentence breaks, spurious sentence breaks, and lost text
1%	confusion of picture captions or section headers with main text
2%	mark-up handling, such as `<math>` blocks and special environments

Parsing Errors	
4%	resource limitations (cpu time or size of search space)
5%	named-entity and other unknown-word handling problems
10%	parse ranking: good analysis likely available, but not in top 500
15%	grammar shortcomings, e.g. pseudopassives (... *is referred to as*)
1%	text errors: typos, missing determiners, *its* vs. *it's* confusion, etc.
1%	miscellaneous: tagger errors, foreign language illustrations, etc.

Table 1: Distribution of coarse-grained error types, including some examples and estimated overall percentages.

with adjacent 'regular' tokens (i.e. non-markup and non-punctuation ones) and then syntactically analyzed as pseudo-affixes. This approach has the benefit of eliminating attachment ambiguities for punctuation and markup tokens (they always attach lowest, i.e. lexically), and furthermore it yields a perfect predictor of standard whitespace conventions around punctuation marks—commas, for example, are pseudo-suffixes; opening parentheses, on the other hand, are pseudo-prefixes. Aligning the treatment of some markup with the existing analysis of punctuation provides a fruitful starting hypothesis for our WeScience experiments. In the case of italicized phrases, the grammar is thus enabled to apply its existing apparatus for 'recognizing' quoted expressions (as an uninterpreted, strictly left-branching binary tree). Once a complete, properly bracketed phrase is recognized, unary rules map the corresponding constituent into a suitable syntactic category, in this case a proper name.

5 Outlook—The Immediate and Mid-Term Future

We embarked on the WeScience effort primarily to enable our own research (on large-scale semantic parsing of scholarly literature, 'deep' information extraction, and ontology learning). But we believe it is worth documenting our results, even in this early stage, because Wikipedia (and other) sub-domain corpora could develop into valuable NLP resources; particularly so, if we can find ways of extending the Wikipedia 'community' approach from creating the original content to the creation of additional linguistic annotation.

In our view, our very preliminary results are encouraging in multiple ways. The combination of an initial seed of documents (in our case derived from the Wikipedia category system, no matter how immature its current status) and simple link cardinality analysis provides a straightforward way of sub-domain extraction. Seeing remaining fuzziness in the notion of our 'target domain' (i.e. NLP-related articles), we see no objective metric to gauge the success of our initial experiment. But in informal presentations to multiple (external) colleagues, we have found our intuition confirmed that our current WeScience collection exhibts a (much) greater degree of domain coherence than the original seed set.

Pre-processing a richly marked-up hypertext for NLP purposes—finding a good balance between human and machine readability, on the one hand, and preserving anything that potentially has bearing on linguistic content, on the other hand—is a challenging exercise in its own right. Through an early public release of the WeScience corpus, we hope to be able to gather community feedback on this aspect of our proposal.

As regards the application of off-the-shelf NLP tools—sentence segmentation, grammatical analysis, and discriminant-based treebanking—and the interactions of document mark-up and linguistic analysis, our simple-minded experiments seem to suggest both a basic level of feasibility and a substantial remaining potential for improvement. As in most (precision-oriented) NLP, we expect that an iterative feedback loop of in-depth error analysis and careful adaptation of the processing pipeline (both at the corpus creation and parsing layers) will facilitate substantial improvements in segmentation, parsing, and treebanking. This work, in itself, will illuminate relevant linguistic properties of Wikipedia-like content, and of course of the open-source NLP resources involved.

Even with substantial improvements in treebanking coverage, say up to the levels reported by Oepen et al. (2004) (of about 85 percent), to actually make the WeScience Treebank useful as a *treebank*, we will need to address the problem of remaining coverage gaps (i.e. out-of-scope inputs from the ERG perspective; be that owed to actual non-grammaticality or lacking grammatical coverage). We expect to adapt the robust parsing approach of Zhang & Kordoni (2008), extend it to facilitate robust meaning composition, and integrate it in the Redwoods environment with a novel facility for post-editing (this is similar to the techniques used in the construction of the Alpino treebank; Bouma et al., 2001).

The success of open-source projects and particularly the Linux operating system is at times attributed to the *release early, release often* paradigm used successfully to coordinate the efforts of distributed communities of developers. In the hope that our WeScience efforts may stimulate adaptation by others, we plan to

release a first version in early 2009.[8] This version will minimally provide a stable selection of in-domain articles and gold-standard sentence segmentation. A complete, treebanked version—though excluding a residual percentage of out-of-scope items—will be made available in the first half of 2009. In joint work with the LinGO team, we expect to adapt both the grammar (extend or improve linguistic analyses) and parsing technology in the light of the WeScience experience.

References

Adolphs, P., Oepen, S., Callmeier, U., Crysmann, B., Flickinger, D., & Kiefer, B. (2008). Some fine points of hybrid natural language parsing. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech, Morocco.

Bouma, G., van Noord, G., & Malouf, R. (2001). Alpino. Wide-coverage computational analysis of Dutch. In W. Daelemans, K. Sima-an, J. Veenstra, & J. Zavrel (Eds.), *Computational linguistics in the Netherlands* (pp. 45 – 59). Amsterdam, The Netherlands: Rodopi.

Carter, D. (1997). The TreeBanker. A tool for supervised training of parsed corpora. In *Proceedings of the Workshop on Computational Environments for Grammar Development and Linguistic Engineering.* Madrid, Spain.

Denoyer, L., & Gallinari, P. (2006). The Wikipedia XML corpus. In *Proceedings of the Fifth Workshop of the Initiative for the Evaluation of XML Retrieval.* Dagstuhl, Germany.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering, 6 (1),* 15 – 28.

Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of the 2001 conference on Empirical Methods in Natural Language Processing.* Pittsburgh, PA.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English. The Penn Treebank. *Computational Linguistics, 19,* 313 – 330.

Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. *Journal of Research on Language and Computation, 2*(4), 575 – 596.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank. A corpus annotated with semantic roles. *Computational Linguistics, 31*(1), 71 – 106.

Rosén, V., De Smedt, K., & Meurer, P. (2006). Towards a toolkit linking treebanking and grammar development. In *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories* (pp. 55 – 66). Prague, Czech Republic.

Rupp, C., Copestake, A., Teufel, S., & Waldron, B. (2007). Flexible interfaces in the application of language technology to an eScience corpus. In *Proceedings of the UK eScience Programme All Hands Meeting.* Nottingham, UK.

Tateisi, Y., Yakushiji, A., Ohta, T., & Tsujii, J. (2005). Syntax annotation for the GENIA corpus. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing* (pp. 222 – 227). Jeju, Korea.

Zhang, Y., & Kordoni, V. (2008). Robust parsing with a large HPSG grammar. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech, Morocco.

[8] This initial release of the corpus and partial treebank will be made available on the WeScience home page: `http://www.delph-in.net/wescience/`.