Towards Universal Web Parsebanks

Juhani Luotolahti[1], **Jenna Kanerva**[1,2], **Veronika Laippala**[3,4]
Sampo Pyysalo[1], **Filip Ginter**[1]
[1]Department of Information Technology
[2]University of Turku Graduate School (UTUGS)
[3] Turku Institute for Advanced Studies, University of Turku, Finland
[4] School of Languages and Translation Studies, University of Turku, Finland
University of Turku, Finland
first.last@utu.fi

Abstract

Recently, there has been great interest both in the development of cross-linguistically applicable annotation schemes and in the application of syntactic parsers at web scale to create parsebanks of online texts. The combination of these two trends to create massive, consistently annotated parsebanks in many languages holds enormous potential for the quantitative study of many linguistic phenomena, but these opportunities have been only partially realized in previous work. In this work, we take a key step toward universal web parsebanks through a single-language case study introducing the first retrainable parser applied to the Universal Dependencies representation and its application to create a Finnish web-scale parsebank. We further integrate this data into an online dependency search system and demonstrate its applicability by showing linguistically motivated search examples and by using the dependency syntax information to analyze the language of the web corpus. We conclude with a discussion of the requirements of extending from this case study on Finnish to create consistently annotated web-scale parsebanks for a large number of languages.

1 Introduction

The enormous potential of the web as a source of material for linguistic research in a wide range of areas is well established (Kilgarriff and Grefenstette, 2003), with many new opportunities created by web-scale resources ranging from simple N-grams (Brants and Franz, 2006) to syntactically analyzed text (Goldberg and Orwant, 2013). Yet, while the use of multilingual web data to support linguistic research is well recognized (Way and Gough, 2003), cross-linguistic efforts involving syntax have so far been hampered by the lack of consistent annotation schemata and difficulties relating to coincidental differences in the syntactic analyses produced by parsers for different languages (Nivre, 2015).

The Universal Dependencies (UD) project[1] seeks to define annotation schemata and guidelines that apply consistently across languages, standardizing e.g. part-of-speech tags, morphological feature sets, dependency relation types, and structural aspects of dependency graphs. The project further aims to create dependency treebanks following these guidelines for many languages. The effort builds on many recently proposed approaches, including Google universal part-of-speech tags (Petrov et al., 2012), the Interset inventory of morphological features (Zeman, 2010) and Universal Stanford Dependencies (de Marneffe et al., 2014), and previously released datasets such as the universal dependency treebanks (McDonald et al., 2013). The first version of UD data, released in early 2015, contains annotations for 10 languages: Czech, English, Finnish, French, German, Hungarian, Irish, Italian, Spanish, and Swedish.

The availability of the UD corpora creates a wealth of new opportunities for the cross-linguistic study of morphology and dependency syntax, which are only now beginning to be explored. One particularly exiting avenue for research involves the combination of these annotated resources with fully retrainable parsers and web-scale texts to create massive, consistently annotated parsebanks for many languages. In this study, we take the first steps toward realizing these opportunities by producing a UD parsebank of Finnish comprising well over 3 billion tokens, and combining it with a scalable query system and web

[1]http://universaldependencies.github.io/docs/

interface, thus building a large-scale corpus and pairing it with the tools necessary for its efficient use. Using real-world examples, we show how the large web corpus with the syntactic annotation can be used for gathering data on rare phenomena in linguistic research.

For linguistic research web corpora, containing broad scope of text, are well suited for the search of rare linguistics constructs as well as those which do not often appear on official text, such as the use of colloquial terms and structures. Other motivations beyond linguistic research for large web-corpora alone are found in natural language processing, for example in language modeling which has uses in many areas such as information extraction and machine translation(Kilgarriff and Grefenstette, 2003).

We finish with a discussion of how to generalize our effort from one language to many, arguing that the framework and tools introduced as one of the primary contributions of this study present many opportunities and can meet the challenges for creating web parsebanks all for all existing UD treebanks.

2 Data

We next briefly introduce the manually annotated corpus used to train the machine learning-based components of our processing pipeline and the sources of unannotated data for creating the web parsebank.

2.1 Annotated data

For training the machine learning methods that form the core of the text segmentation, morphological analysis, and syntactic analysis stages of the parser, we use the Universal Dependencies (UD) release 1.0 Finnish corpus (Nivre et al., 2015). This corpus was created by converting the annotations of the Turku Dependency Treebank (TDT) corpus (Haverinen et al., 2014) from its original Stanford Dependencies (SD) scheme into the UD scheme using a combination of automatically implemented mapping heuristics and manual revisions. TDT consists of documents from 10 different domains, ranging from legal texts and EU parliamentary proceedings, through Wikipedia and online news to student magazine texts and blogs. In total, the UD Finnish data consists of 202,085 tokens in 15,136 sentences, making it a mid-sized corpus among the ten UD release 1 corpora, which range in size from 24,000 tokens (Irish) (Lynn et al., 2014) to over 1,5 million tokens (Czech) (Bejček et al., 2012).

2.2 Unannotated data

We use two web-scale sources of unannotated text data: the openly accessible Common Crawl dataset,[2] and data produced by our own large-scale web-crawl, introduced in Section 3.1. Common Crawl is a non-profit organization dedicated to producing a freely available reference web crawl dataset of the same name. As of this writing, the Common Crawl consists of several petabytes (10^{15}) of data collected over a span of 7 years, available through the Amazon web services Public Data Sets program.[3]

While web datasets such as the Common Crawl represent enormous opportunities for linguistic efforts, it should be noted that are many known challenges to extracting clean text consisting of sentences with usable syntactic structure from such data. For one, text content must primarily be extracted from HTML documents, and thus contains many lists, menus and other similar elements not (necessarily) relevant to syntactic analysis. Indeed, such text not consisting of parseable sentences represents the majority of all available text (see Section 4.1), necessitating a filtering step. Another major issue is the large prevalence of duplicate content due to advertisements often appearing on many domains, many sites hosting copied content, such as the contents of the Wikipedia, in order to generate traffic and search engine hits, and sites such as web forums containing many URLs with overlapping content (e.g. URLs which highlight a specific comment of the thread). We discuss the ways in which we address these issues in the following section.

3 Methods

In the following, we present the primary processing stages for building the parsebank, summarized in Figure 1, and the search system used to query the completed parsebank.

3.1 Dedicated web crawl

The currently existing non-UD Finnish Internet parsebank (Kanerva et al., 2014) is based on texts

[2] http://commoncrawl.org/

[3] http://aws.amazon.com/public-data-sets/

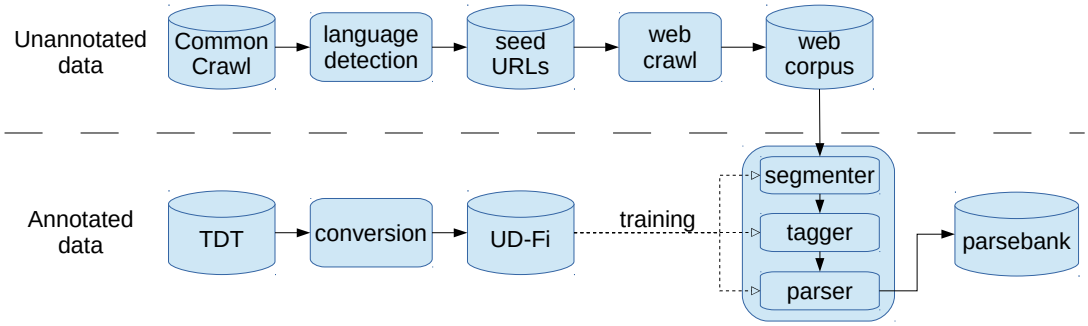

Figure 1: Processing stages. Seed URLs are first selected from Common Crawl data using language detection, and a web crawl is then performed using these seeds to identify an unannotated web corpus. To train the text segmentation, morphological tagging, and parsing stages of the analysis pipeline, UD Finnish data created by semiautomatic conversion of Turku Dependency Treebank is used. The final web parsebank is then created by applying the trained analysis pipeline on the unannotated web corpus.

extracted from the 2012 release of the Common Crawl dataset using the Compact Language Detector.[4] This 1.5 billion token corpus was assembled from approximately 4 million URLs. However, as this dataset based solely on Common Crawl data fell somewhat short of our target corpus size, we expand it as part of this study with a dedicated crawl targeting Finnish.

To seed the crawl, we obtained all public domains registered in the Finnish top level domain (`.fi`) and extracted all the URLs from the current Common Crawl-based Finnish Internet parsebank. This allows us to reach as wide a scope as possible, going beyond the Finnish top-level domain. Following the identification of the seed URLs, the final web corpus data used to build the parsebank was crawled using the open source web crawler SpiderLing (Suchomel and Pomikálek, 2012). SpiderLing is designed for collecting unilingual text corpora from the web. During the crawl, the language of each downloaded page is recognized to maintain the language focus of the crawl. The language recognition, a built-in feature of the crawler, is based on character trigrams. Similarly, the character encoding of the content is heuristically determined during the processing, and allows the content to be encoded into the standard UTF-8 encoding when storing the data for further processing.

Supporting a focus on text-rich pages, SpiderLing also keeps track of the text yield of each domain, defined as the amount of text gathered from a domain divided by the amount of bytes downloaded, and prioritizes domains from which can be obtained more usable data in less time. The crawler also makes an effort to gather only text content from the web, avoiding downloading other content such as images, javascript, etc. Further, to extract clean text consisting of sentences, as opposed to lists, menus and the like, the crawler automatically performs boilerplate removal, using the `justText` library. The usable text detection is based on various metrics such as the frequency of stop words in a given paragraph, link density, and the presence of HTML-tags. (Text deemed as boilerplate is ignored when calculating the yield.)

The crawl was performed on a single server-grade Linux computer in a series of bursts between the summer and winter of 2014, taking approximately 88 days. The crawl speed settings were kept very conservative to prevent causing false alarms to Internet security authorities. The text data from the old corpus will be merged in the corpus, but for now the result of this crawl is the source for all text in this version of the web corpus.

3.2 Text segmentation

For the segmentation of raw text into sentences and then further into tokens, we apply the machine-learning based sentence splitter and tokenizer from the Apache OpenNLP toolkit[5]. Both the sentence splitter and the tokenizer are retrainable maximum entropy-based systems, and we trained new models for both based on the data from the UD Finnish corpus.

[4] `https://code.google.com/p/cld2/`

[5] `https://opennlp.apache.org/`

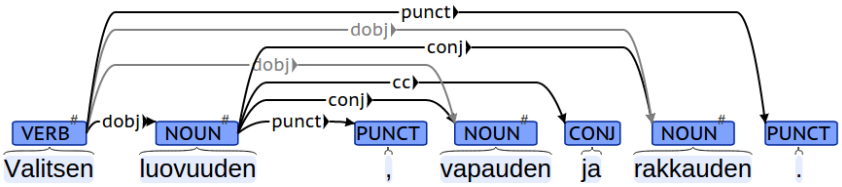

Figure 2: An example UD analysis of a Finnish sentence *Valitsen luovuuden, vapauden ja rakkauden* "I_choose creativity, freedom, and love." Extended dependencies produced by propagating the object dependency into the coordinated constituents are shown in gray. Figure created using BRAT (Stenetorp et al., 2012).

3.3 Morphological tagging

To assign the part-of-speech tags and the morphological features to words, we apply the Conditional Random Fields (CRF)-based tagger Marmot (Mueller et al., 2013), deriving lemmas and supplementing the feature set of the retrainable tagger with information derived from a pipeline combining the finite-state morphological analyzer OMorFi (Pirinen, 2011) with previously introduced heuristic rules for mapping its tags and features into UD (Pyysalo et al., 2015).

Our previous evaluation of the morphological analysis components on the UD Finnish data indicated that the best-performing combination of information derived from the finite-state analysis and the machine learning system allowed POS tags to be assigned with an accuracy of 97.0%, POS tags and the full feature representation with an accuracy of 94.0%, and the complete morphological analysis, including the lemma, with an accuracy of 90.7% (Pyysalo et al., 2015). This level of performance represents the state of the art for the analysis for Finnish and is broadly comparable to the state-of-the-art results for these tasks in other languages.

3.4 Syntactic analysis

The dependency parsing is carried out using the graph-based parser of Bohnet et al. (2010) from the Mate tools package, trained on the UD Finnish data. The parser has previously been evaluated on the test section of the TDT corpus, achieving 81.4% LAS (labeled attachment score). This approaches the best test score of 83.1% LAS reported in the study of Bohnet et al. (2013) using a parser that carries out tagging and dependency parsing jointly.[6] However, at approximately 10ms per sentence, the graph-based parser is an order of magnitude faster than the more accurate joint tagger and parser, which is a deciding factor when parsing billions of tokens of text. When re-training the graph-based parser on the UD scheme annotations, it achieved a LAS of 82.1% on the UD Finnish test set, showing that the parsing performance is not in any way degraded compared to that for the original SD scheme of the treebank.

In addition to the *basic* layer of dependencies, which constitutes dependencies that form a tree structure, the parsing pipeline also predicts the UD Finnish *extended* layer dependencies, modeled after the conjunction propagation and external subject prediction in the original SD scheme (de Marneffe and Manning, 2008). This layer anticipates the introduction of such an extended layer into the UD scheme, which allows additional, non-tree dependencies in terms of its format but only presently provides guidelines for the *basic* layer. The extended layer prediction is based on the method of Nyblom et al. (2013), originally developed on the TDT corpus SD scheme, re-trained and adapted for the current study to conform to the UD scheme. An example parse with extended layer dependencies is shown in Figure 2.

3.5 Parsebank search

A parsebank of the billion token magnitude is only useful if it can be efficiently queried, especially taking advantage of the syntactic structures, i.e. using queries which would be difficult or impossible to express in terms of the linear order of the words. We have therefore previously developed a scalable syntactic structure query system which can be applied at this scale and allows rich syntactic structure queries referring to both the basic

[6] Note that results are for the original SD annotation of the TDT corpus. While the UD Finnish treebank is created from this data (primarily) by deterministic conversion, the results are thus not fully comparable with results for the UD Finnish corpus.

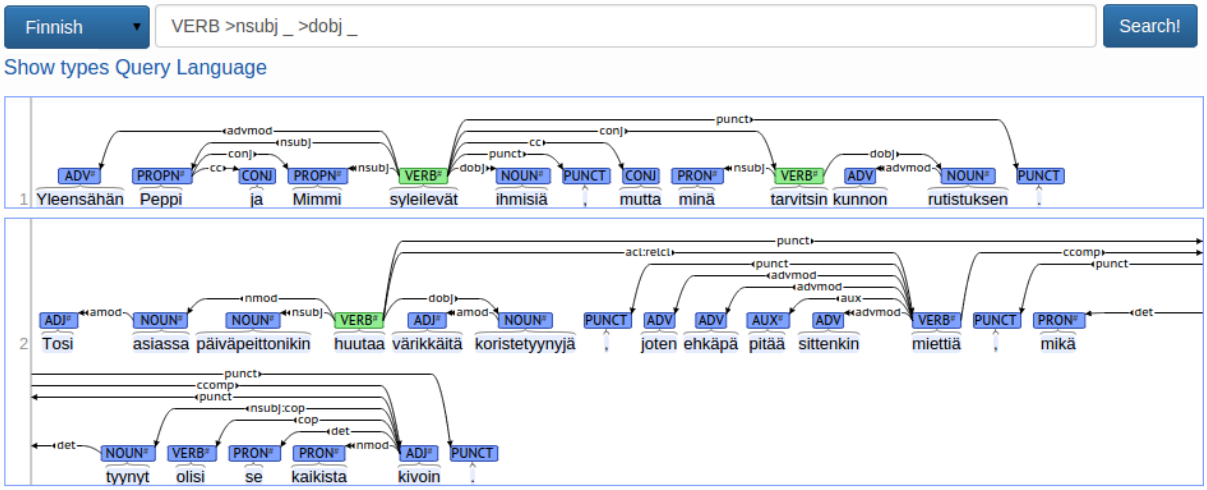

Figure 3: A screenshot of the online query interface, showing a simple query for transitive verbs.

and the extended layers of the analysis (Luotolahti et al., 2015). This detailed corpus search enables fast and easy retrieval of material for many linguistic questions that otherwise would require manual work to address.

The query system allows search for any arbitrary subtree structure, including arbitrarily nested negations. For instance, one can search for verbs which have their subject in the partitive case, unless that subject has a numeral modifier, and unless the verb is governed by the clausal complement relation. In addition to the constraints on the syntactic structure, any combination of normal and negated constraints on the morphology of the words is possible. The full description of the query system capabilities is, however, out of scope of this paper, and we refer the interested reader to the online documentation [7]. In addition to a scriptable, command-line utility meant for gathering data for further processing, the query system also has an online interface which allows the results to be visualized and inspected in real time (Figure 3).

In Section 5 we will demonstrate several real use-cases where this query system was used to obtain material for linguistic research from the parsebank.

4 Results

We next briefly present the primary quantitative results of our study, the web corpus created as the result of our custom crawl, the performance characteristics of our newly trained parsing pipeline,

[7] http://bionlp.utu.fi/searchexpressions-new.html

Item	Number
All tokens	3,662,727,698
Lemma count	28,585,422
Sentence count	275,690,022
Unique token count	39,688,642
Unique sentence count	178,547,962
Tokens without duplicates	2,554,094,599

Table 1: Web corpus statistics.

Item	Number
All tokens	94,528,120
Lemma count	1,532,485
Sentence count	8,477,560
Unique token count	3,067,151
Unique sentence count	7,252,240
Tokens without duplicates	87,772,532

Table 2: News data statistics.

and some statistical characteristics of the web corpus. For reference, we contrast the web corpus to the *news* section of the Finnish Text Collection (*Suomen kielen tekstikokoelma*) corpus[8], below referred to as the news corpus, as these news domain texts are a typical representative of a conventional corpus used for linguistic research.

4.1 Web crawl results

The web crawl retrieved in total 1.6 terabytes of HTML pages over the 88 days it was run. From this HTML data, 170 gigabytes of plain prose text was extracted, excluding markup and boilerplate content such as menus. This body of text still con-

[8] http://urn.fi/urn:nbn:fi:lb-201403268

tained a significant amount of duplication, which was eliminated on the document level in order to preserve the document context of the sentences in the parsebank. The deduplication process determined a document as a duplicate if more than 90% of its sentences were seen earlier during a sweep through the data. Following this deduplication process, the resulting final web corpus is 33 gigabytes in size, i.e. only approximately 2% of the total data downloaded by the crawler. The basic statistics of the resulting corpus are given in Table 1, and corresponding statistics for the news corpus are presented in Table 2. We note that the web corpus is an order of magnitude or more larger that the extensive newswire corpus by any metric, most notably containing nearly 40 times the number of tokens of the conventional dataset.

4.2 Parsing accuracy and speed

The syntactic parsing pipeline has previously only been evaluated on the test set of the UD Finnish dataset, which closely reflects the distribution of the training data in terms of topics, genres and styles of writing. On this test set, the parser achieved 82.1% LAS on the basic UD dependencies. To evaluate how well the parser generalizes to out-of-domain web data, we selected a random 100 sentences from the parsebank and manually annotated them for UD syntax (both basic and extended layers). In the process, we discarded two incomprehensible sentences, most likely produced by a machine translation system, for which it was not possible to arrive at a reasonable gold standard tree. We were then left with 98 sentences comprising 1,191 tokens. On this sample, the LAS of the parsing pipeline is 78.1% when we take the extended layer into account (a token is counted as correct if it is correctly attached in both the basic and extended layers), and 78.8% for the basic layer only. This about 3% point drop (from 82.1% to 78.8% LAS on UD basic layer) is quite acceptable considering that the parser has not been adapted to the general web text domain in any way. Dependency parsing errors of an earlier iteration of the same parsing-pipeline for Finnish using very related SD-scheme are analyzed in-depth by Haverinen et al.(2011).

The parsing was carried out on a cluster computer comprising thousands of compute nodes, and took approximately 8,000 CPU core hours (roughly one CPU-year), which due to the highly parallel nature of the process was completed in a little over one day. While parsing is the most computationally demanding component of the overall process of creating the parsebank, it is thus not likely to be a bottleneck for real-time work in generalizing to other languages, even if web corpora of an order of magnitude larger were considered.

4.3 Web corpus characteristics

In corpus linguistics, a standard method to provide an overview of corpus contents is offered by *keyword analysis* (Scott and Tribble, 2006). Describing statistically the most typical words of the studied corpus in relation to a reference one, keywords are typically informative on the corpus theme and style. Table 3 presents keywords extracted from the entire web corpus together with those for the news corpus used for reference. The keywords are calculated using the most significant text class features assigned to the two corpora by a linear classifier trained to distinguish short segments of the two corpora.[9] The classifier is trained using the stochastic gradient method, with a 50/50 split on testing and training data, using labeled text segments five sentences long.

The keywords presented are based on the 50 most significant tokens for the parsebank and 30 for the News corpus. Individual characters and figures are excluded from the table. As can be seen from the number of keywords presented, this is already revealing: numbers and individual characters are clearly more frequent features in the parsebank than in the news text. The actual keywords listed reflect the characteristic topics in the two corpora. The parsebank keywords include terms related to online stores, TV shows and social media. In particular the emoticon is a typical example of computer-mediated text. The news corpora keywords, in contrast, are mainly composed of the names of Finnish towns, political parties and news agencies. An interesting detail is the apparition of the former Finnish currency (*markka*, used until 2001) on the list. This is explained by the fact that the new corpus dates from the 1990s; in the more recent Finnish Internet parsebank, this old form of currency is obviously referred to considerably less frequently.

[9] Implemented using the Vowpal Wabbit machine learning package (Agarwal et al., 2014)

Parsebank keywords
euroa, lue, sosiaali-, ;), vs, tuotantokausi, työ, yms, 1990-luvun, eurolla, kommentit, kommenttia, tiivistelmä, voit, blogissa, blogi
Parsebank keywords in English
euros, read, social-. ;), vs, season (as in TV shows), work, etc, of-the-1990s, with-a-euro, comments, a-comment, summary, you-can, in-a-blog. a-blog
News Corpus keywords
karjalaisen, aamulehden, luvulla, kosovon, reuters, lieksan, tv, hhh, markalla, pohjois-karjalassa, ws, lehtikuva, n., demarin, pohjois-karjalan, joensuussa, joensuu, markan, joensuun, markkaa, demari, stt
News Corpus keywords in English
from-carelia (Finnish region), of-aamulehti (newspaper), with-the-figure, of-kosovo, reuters, of-lieksa (town), tv, hhh, with-a-mark, in-northern-carelia, ws, lehtikuva (Finland's leading photo agency), about (abbreviation), of-a-social-democrat (colloquial), of-northern-carelia, in-joensuu (town), joensuu, of-mark, of-joensuu, marks, social-democrat (colloquial), stt (abbreviation of a Finnish media outlet)

Table 3: Keywords of the parsebank texts in comparison with the news corpus.

5 Linguistic applications

We next illustrate the applicability of the web parsebank and the search system through three linguistically motivated applications based on real-world use-cases.

Web corpora with dependency syntax analyses can considerably speed-up the material collection in research of extremely rare phenomena, here exemplified by Finnish transitive sentences with a partitive subject (Huumo, 2015). Being unnormative, they cannot be easily found from edited or professionally written texts, which also makes web-crawled data a very convenient source for these constructions. In addition, gathering these examples from large corpora without the support of syntactic analyses would be extremely time-consuming. Unfortunately, the rarity of the construction also causes problems in the accuracy of their syntactic analysis. For instance, the parser training data does not have even a single example, and the parser thus tends to make errors in the analysis of this construction, often swapping the subject and the object of the verb (in Finnish, both the subject and the object can take the partitive case). In practice, when listing a random sample of candidate occurrences for manual inspection, the vast majority of these will be incorrect. Nevertheless, even though correct instances are rare in the parsebank, the speed-up in gathering real examples is enormous, considering the alternatives. To illustrate this, we consider the verb *seurata* "to follow" which is theoretized to be especially susceptible for this use. In a sample of 4 million sentences, we find 7,875 transitive occurrences of the verb, of which 111 have their subject in the partitive case, and of these 13 are correct. While this fraction is small, manually inspecting the roughly 100 occurrences took little effort and resulted in real examples being found from among a large number of occurrences of the verb.

Query	Results
koska "because" + no verb	22598
koska "because" + verb	505514

Table 4: Example queries and their results.

Conjunction	Occurrences
ja "and"	738372
mutta "but"	533683
eli "or", "in other words"	153180
tai "or"	110639
vaan "but"	9908
mut / "but" (colloquial)	25057
Total	1671041

Table 5: Sentence-initial conjunction frequencies.

Another example of a construction for which a web-based, syntactically analyzed corpus is very convenient is the new usage of the Finnish subordinating conjunction *koska* "because" (Sinnemaa, 2014; Rehn, 2014). Normatively, a subordinating conjunction should be used in an subordinate clause with a finite verb, attached to the main clause, *I ate because I was hungry*. However, Finnish has recently seen a construction where the subordinate clause is left without the finite verb, but the conjunction is still present, in particular in informal language varieties: *I ate because hungry*. Since this construction is relatively infrequent, traditional corpora without syntactic information can not be used to study the phenomenon. The syntactic analyses in the parsebank, however, enable the search for this exact construction. Table 4 shows the results of a search for *koska* "because" governed by a verb and governed by a noun. As can be seen, although the normative usage with a verb is much more frequent, the search retrieves also a useful number of occurrences where the conjunction is attached to a noun.

Finally, although the automatic analyses only concern syntax and morphology, they can also

be used to retrieve material to study phenomena crossing the limits of individual sentences, such as semantic relations between text elements and discourse structure (Prasad et al., 2008; Laippala et al., 2015). As the search tool allows the restriction of the query to certain sentence elements, it can be delimited to sentence-initial elements, such as sentence-initial, individual conjunctions that instead of co-ordinating sentence-internal clauses or phrases refer to previous text elements and express relations between sentences and the discourse structure. This can provide useful information both on the frequency of different conjunctions used in this position and on discourse structure more in general. The distribution of the most frequently used conjunctions in this functions is presented in Table 5. The results conform to expectations, with *and* being the most frequent conjunction. The frequency of the colloquial form of *but* also illustrates the nature of the parsebank text.

6 Discussion

We have demonstrated the feasibility of creating a UD web parsebank at the scale of billions of words and making it searchable for complex syntactic patterns. However, our efforts in this study have a very obvious limitation, namely only involving a single language. To realize the full potential of web-scale parsebanks annotated using the cross-linguistically consistent UD scheme, this work must be extended to cover several languages, preferably at least the ten languages covered in the current, first release of UD data. We next briefly consider the technical requirements and computational costs of this extension.

First, the parsing pipeline applied here should be largely straightforwardly applicable to currently available UD languages. The core segmentation, morphological analysis, and dependency parsing components of the parser are all fully retrainable, and each implemented using approaches that achieve levels of performance broadly comparable with the state of the art for their respective tasks in the ten UD release 1 languages. A minor issue is the lack of finite-state morphological analyzers (comparable to OMorFi here) for many of the languages, but previous results suggest that the benefits of such a component may be modest for other UD languages, which are generally not as morphologically complex as Finnish (Bohnet et al., 2013). We anticipate that different strategies to tokenization will eventually become necessary to generalize the approach to languages written using systems that do not involve white-space token boundaries, such as Japanese and Chinese. However, no such language is included in the initial set of UD languages.

Second, the language considered in this case study, Finnish, is comparatively rare on the web compared to most of the UD languages. This can be considered both a positive and a negative for generalization to other languages. On the positive side, it is much easier to create corpora of comparable size (billions of tokens) for languages such as English, French, German and Spanish. Indeed, Common Crawl data will suffice, removing the need to extend the data with a custom crawl. However, it is considerably more challenging to create web corpora for such languages that would represent a substantial fraction of the web in that language, and even if such a web corpus were available, the computational cost of parsing it could become infeasible for the somewhat limited resources at our disposal. For these reasons, we will limit our near future efforts in creating the first set of universal web treebanks to similar scale as here for all considered languages (or smaller when not available for a language). We will also primarily rely on Common Crawl data, only performing additional crawls when this data fails to meet the target size for a language.

As there are no components in the processing pipeline that would scale more than linearly in their computational cost with respect to the number of sentences and we will not aim to substantially increase the size of any language-specific corpus over that created here, we expect the total computational cost of scaling from one language to ten to be simply an order of magnitude greater than that here. Thus, we estimate that the total computational cost of creating the first set of UD web parsebanks to be on the order of 100,000 CPU core hours. While this is a non-trivial cost, it is well within our resources.

7 Conclusions and future work

We have proposed to create universal web parsebanks, web-scale corpora in many languages that are automatically syntactically analyzed using the cross-linguistically consistent Universal Dependencies (UD) scheme. We have also taken a key step toward realizing this possibility in building a

UD Finnish parsebank as a case study. Seeding a web crawl from Common Crawl data, we created the largest Finnish Internet language web corpus of over 3 billion tokens, trained a state-of-the-art dependency parser on the manual UD Finnish corpus annotation, and applied the trained parser to produce the first UD parsebank. We then demonstrated the application of the parsebank to linguistically motivated tasks by integrating it into a scalable dependency corpus search system and supporting several real-world use cases focusing on the identification of relevant examples of rare phenomena.

In future work, we will extend this effort to cover all ten of the UD release 1.0 languages – Czech, English, Finnish, French, German, Hungarian, Irish, Italian, Spanish, and Swedish – to create the first set of cross-linguistically consistently annotated web treebanks, which will be made freely available under open licenses.

Acknowledgments

This work was supported by the Kone Foundation and the Emil Aaltonen Foundation. Computational resources were provided by CSC – IT Center for Science. Data from the Common Crawl foundation was used for web crawling.

References

Alekh Agarwal, Olivier Chapelle, Miroslav Dudi, and John Langford. 2014. A reliable effective terascale linear learning system. *JMLR*, 15:1111–1133.

Eduard Bejček, Jarmila Panevová, Jan Popelka, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, and Zdeněk Žabokrtskỳ. 2012. Prague dependency treebank 2.5 – a revisited version of pdt 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING'10*, pages 89–97.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. Linguistic Data Consortium. LDC2006T13.

Marie-Catherine de Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University. http://nlp.stanford.edu/software/dependencies manual.pdf.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, volume 14, pages 4585–4592.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247.

Katri Haverinen, Filip Ginter, Veronika Laippala, Samuel Kohonen, Timo Viljanen, Jenna Nyblom, and Tapio Salakoski. 2011. A dependency-based analysis of treebank annotation errors. In *Proceedings of International Conference on Dependency Linguistics (Depling'11), Barcelona, Spain*, pages 115–124.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3):493–531.

Tuomas Huumo. 2015. The partitive A: On uses of the Finnish partitive subject in transitive clauses. In *Diachronic typology of differential argument marking*.

Jenna Kanerva, Juhani Luotolahti, Veronika Laippala, and Filip Ginter. 2014. Syntactic n-gram collection from a large-scale corpus of internet finnish. In *Proceedings of the Sixth International Conference Baltic HLT*, pages 184–191.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

Veronika Laippala, Jenna Kanerva, Anna Missilä, Katri Haverinen, Tapio Salakoski, and Filip Ginter. 2015. Towards a discourse-annotated corpus of finnish: the finnish propbank. In *Poster presented at the TextLink Cost Action seminar, Louvain-la-Neuve, Belgium, 27.1.2015*.

Juhani Luotolahti, Jenna Kanerva, Sampo Pyysalo, and Filip Ginter. 2015. Sets: Scalable and efficient tree search in dependency graphs. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 51–55.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 92–97.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.0. Available: http://hdl.handle.net/11234/1-1464.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Jenna Nyblom, Samuel Kohonen, Katri Haverinen, Tapio Salakoski, and Filip Ginter. 2013. Predicting conjunct propagation and other extended stanford dependencies. In *Proceedings of the International Conference on Dependency Linguistics (Depling 2013)*, pages 252–261.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, pages 2089–2096.

Tommi A Pirinen. 2011. Modularisation of Finnish finite-state language description–towards wide collaboration in open source development of a morphological analyser. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 299–302.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (Nodalida 2015)*, pages 163–173.

Anneliise Rehn. 2014. Because meaning: Language change through iconicity in internet speak. In *2014 SURF Conference Proceedings*.

Mike Scott and Christopher Tribble. 2006. *Textual patterns: key words and corpus analysis in language education*. John Benjamins.

Tiina Sinnemaa. 2014. Ei saa ronkkia ruokaa, koska afrikan lapset! koska np-rakenteen merkityksest ja ilmaisuvoimasta / you should not play with your food because [of] the children in africa! on the significance and expressivity of the construction 'because np'. Bachelor's thesis, Department of Finnish, University of Turku.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Vít Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

Andy Way and Nano Gough. 2003. webmt: developing and validating an example-based machine translation system using the world wide web. *Computational Linguistics*, 29(3):421–457.

Daniel Zeman. 2010. Hard problems of tagset conversion. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pages 181–185.