The Dundee Treebank

Maria Barrett, Željko Agić and Anders Søgaard

Center for Language Technology, University of Copenhagen
Njalsgade 140, 2300 Copenhagen S, Denmark
E-mail:{barrett|zeljko.agic|soegaard}@hum.ku.dk

Abstract

We introduce the Dundee Treebank, a Universal Dependencies-style syntactic annotation layer on top of the English side of the Dundee Corpus. As the Dundee Corpus is an important resource for conducting large-scale psycholinguistic research, we aim at facilitating further research in the field by replacing automatic parses with manually assigned syntax. We report on constructing the treebank, performing parsing experiments, as well as replicating a broad-scale psycholinguistic study—now for the first time using manually assigned syntactic dependencies.

1 Introduction

The *Dundee Corpus* is a major resource for studies of linguistic processing through eye movements. It is a famous resource in psycholinguistics, and—to the best of our knowledge—the world's largest eye-movement corpus. The English part of the Dundee Corpus was annotated with part-of-speech (POS) information in 2009 [9]. This layer of annotation facilitated new psycholinguistic studies such as testing several reader models using models of hierarchical phrase structure and sequential structure [10].

In this paper, we describe a recent annotation effort to add a layer of dependency syntax on top of the POS annotation, enabling the replication of classic studies such as [8] on manually assigned syntax rather than automatic parses. We first describe the Dundee Corpus, then our annotation scheme, and finally we discuss applications of this annotation effort.

2 The Dundee Corpus

The Dundee Corpus was developed by Alan Kennedy and Joël Pynte in 2003, and it contains eye movement data on top of English and French text [13]. Measurements were taken while participants read newspaper articles from *The Independent* (English) or *Le Monde* (French). Ten native English-speaking subjects participated

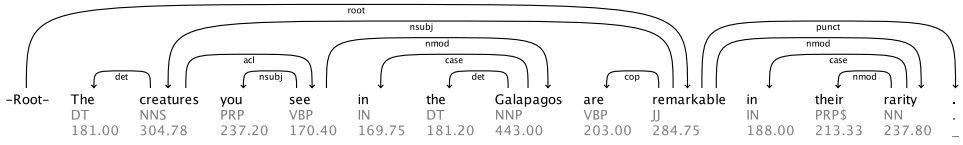

Figure 1: An example sentence (#10) from the Dundee Corpus with UD-style syntactic dependencies and per-word fixation durations.

in the English experiments reading 20 articles, which we focus on here. For a more detailed account, see [14].

The English corpus contains 51,502 tokens[1] and 9,776 types in 2,368 sentences. The apparatus was a Dr Bouis Oculometer Eyetracker with a 1000 Hz monocular (right) sampling. The corpus provides information on fixation durations and fixation order on word level—while also accounting for landing position—for a relatively natural reading scenario. Subjects read running text, 5 lines per display.

Eye movements provide a window to the workings of the brain, e.g. by reflecting cognitive load. Recordings of eye movements during reading is one of the main methods for getting a millisecond to millisecond record of human cognition. Eye movements during reading is controlled by a complex interplay between low-level factors (how much the eye can see and encode from each fixation, word length, landing position, etc.) and high-level factors (e.g. syntactic processing). For an overview, see [18].

This resource has enabled researchers to study things like syntactic and semantic factors in processing difficulty of words [16] and whether the linguistic processing associated with a word can proceed before the word is uniquely identified [19].

3 Syntactic Annotation

In annotating the Dundee Corpus for syntactic dependencies, we follow the Universal Dependencies (UD) guidelines[2] [1] as the emerging *de facto* standard for dependency annotation.

The guidelines build on—and closely adhere to—Universal Stanford dependencies [7], proposing 40 dependency relations together with an universal POS tagset (UPOS) and morphological features. We convert the Penn Treebank-style POS tags from the Dundee Corpus into UPOS, and we provide the universal morphology features, by using the official English UD conversion tools.

The guidelines for annotating English are very well-documented within the UD framework. We only briefly touch upon the most important ones.

For core dependents of clausal predicates, UD distinguishes between nominal subjects (NSUBJ), nominal subjects of passives (NSUBJPASS), direct objects

[1] According to the tokenisation of the Dundee corpus where punctuation and contracted words are glued to the preceding word.

[2] `http://universaldependencies.github.io/`

	Dundee			English UD dev			English UD test		
Training set	*LAS*	*UAS*	*LA*	*LAS*	*UAS*	*LA*	*LAS*	*UAS*	*LA*
Dundee	82.23*	85.06*	89.97*	69.50	75.96	81.26	68.86	75.60	80.61
English UD	71.45	78.66	84.28	85.51	88.03	92.91	84.72	87.30	92.37

Table 1: Dependency parsing results with English UD and Dundee as training sets. Parser: `mate-tools` graph-based parser with default settings [5]. Features: FORM and CPOSTAG only, using the Penn Treebank POS tags. Metrics: labeled and unlabeled attachment scores (LAS, UAS), and label assignment (LA). *: 5-fold 80:20 cross-validation, as the Dundee Treebank has no held-out test set.

(DOBJ), indirect objects (IOBJ), clausal subjects (CSUBJ), clausal subjects of passives (CSUBJPASS), clausal complements (CCOMP), and open clausal complements (XCOMP). When it comes to non-nominal modifiers of nouns, for example, the guidelines distinguishes between adjectival modifiers (AMOD), determiners (DET), and negation (NEG).

We show an example sentence from the treebank in Figure 1. It depicts the UD-style dependency annotation, as well as per-word total fixation durations averaged over ten readers. Some of the typical UD-style conventions—such as content head primacy and no copula heads—are also illustrated.

We used two professional annotators that had previously worked on treebanks following the UD guidelines. The annotators provided double annotations for 118 sentences, with moderately high inter-annotator agreements of 80.82 (LAS), 87.61 (UAS), and 86.63 (LA).

Further, we trained a graph-based dependency parser [5] on English UD training data, and parsed the Dundee Corpus text. We report the results in Table 1. There is a decrease in accuracy moving from English UD to the Dundee Corpus text. We attribute the decrease to the domain shift—English UD stemming from various web sources, while Dundee consists of newswire commentaries in specific—and possibly to the slight cross-dataset inconsistency in POS and dependency annotations. In a separate experiment, we also parse the Dundee Corpus text using 5-fold cross-validation with an 80:20 split, observing accuracies consistent with the English UD experiment. These results are also reported in Table 1.

The cross-dataset decrease in parsing accuracy, even if irrelevant for Dundee-specific experiments, plays into the argument for using gold-standard annotations in psycholinguistic research.

4 Replication of Dependency Locality Theory Experiment

The Dundee Treebank annotated with dependencies has the following affordances. First, it allows for replication of studies such as [8] with manual annotations. Sec-

Predictor	Coef	p	Coef original	p original
INTERCEPT	199.59		128.24	***
WORDLENGTH	-1.25		30.90	***
WORDFREQUENCY	4.43	***	14.50	***
PREVIOUSWORDFIXATED	-33.32	***	-18.05	***
LANDINGPOSITION	-1.23	***	-4.18	***
LAUNCHDISTANCE	1.79	***	-1.91	***
SENTENCEPOSITION	-.09	*	-.12	*
FORWARDTRANSITIONALPROBABILITY	1.51	***	-3.27	***
BACKWARDTRANSITIONALPROBABILITY	-5.87	***	3.96	***
log(DLT)	3.51	**	5.86	*
WORDLENGTH:WORDFREQUENCY	-2.96	***	-4.98	***
WORDLENGTH:LANDINGPOSITION	-.68	***	-1.02	***

Table 2: First pass durations for nouns with non-zero DLT score in the Dundee corpus. Coefficients and their significance level. Same predictors as original noun experiment. * $p < .05$, ** $p < .01$, *** $p < .001$.

ond, gaze features can be used to improve NLP models by enabling joint learning of gaze and syntactic dependencies [2, 3]. Finally, the Dundee Treebank facilitates for researchers to study the reading of very specific syntactic constructions in naturalistic, contextualized text, while controlling for individual variation, and variation specific to the parts of speech or syntactic dependencies involved.

Demberg and Keller(D&K) were the first to test broad-covering theories of sentence processing on large-scale, contextualized text with eye tracking data [8]. They explored two theories of syntactic complexity, namely Dependency Locality Theory (DLT) and Surprisal, and how these correlate with three eye tracking measures while controlling for oculomotor and low-level processing.

DLT [11] estimates the computational resources consumed by the human processor and computes a cost for any discourse referent as well as a cost for every discourse referent between a particular discourse referent and it's head. Thus, DLT needs dependency parsed text to score the complexity of the sentences and Minipar was used to parse the text with a reported 83% accuracy of the DLT score.

In this paper we replicate the parts of their experiments involving DLT, but with manually assigned dependencies instead of automatic parses for calculating DLT. D&K found that DLT score did not have the expected positive effect on reading time of all words. The calculation of DLT only applies for nouns and verbs. They did, however, find that DLT significantly had a positive effect on reading times for nouns and verbs.

We replicate the linear mixed-effects experiment using first pass fixation duration per word for all words and nouns[3]. First pass fixation duration is the duration

[3]The original paper does not contain information about the fixed effects of the model for verbs, why this part of the experiment was not replicated.

of all fixations on specific word from the readers eyes first enter into the region and until the eyes leave the region, given that this region is fixated. This is an measure said to encompass early syntactic and semantic processing as well as lexical access. We use the same low-level predictor variables as the original experiment:

1. word length in characters (WORDLENGTH),
2. log-transformed frequency of target word (WORDFREQUENCY),
3. log-transformed frequency of previous word (PREVIOUSWORDFREQUENCY),
4. forward-transitional probability (FORWARDTRANSITIONALPROBABILITY),
5. backward transitional probability (BACKWARDTRANSITIONALPROBABILITY),
6. word position in sentence (SENTENCEPOSITION),
7. whether the previous word was fixated or not (PREVIOUSWORDFIXATED),
8. launch distance of the fixation in characters (LAUNCHDISTANCE),
9. and fixation landing position (LANDINGPOSITION).

Backward- and forward transitional probabilities are conditional probabilities of a word given the previous / next word, respectively [15]. Along with the word frequencies these two measures are obtained from the British National Corpus (BNC) [6], following the line of D&K. We use KenLM [12] for getting the bigram frequencies and Kneser-Ney smoothing for those bigrams that are not found in the training set. D&K used CMU-Cambridge Language Modeling Toolkit and applied Witten-Bell smoothing. Bigrams respect sentence boundaries.

We clean the data following the described approach by using only fixated words, excluding words that are followed by any kind of punctuation and excluding first and last words of each line. We did, however, not remove words "in a region of 4 or more adjacent words that had not been fixated", since it is unclear what a "region" is (non-fixated words are already removed). This left us with 209,010 datapoints. D&K report to have 200,684 datapoints after cleaning. The difference is probably accounted for by the missing, last cleaning step.

We use R [17] and lme4 [4] to fit a linear mixed-effects model. In the following we use the same fixed and random effects as their models minimised using Akaike Information Criterion (AIC). The authors do not report which significance test they used. We use likelihood ratio tests of the full model with the particular fixed effect against the model without the particular fixed effect.

D&K find that for all words, DLT had a significant, negative effect on first pass fixation duration ($p < .001$), which is a displeasing counter-intuitive result. It means higher DLT score gives a shorter fixation duration. We also find a very small negative effect (-.03) of DLT on first pass fixation duration for all words, but it doesn't reach significance. Following the original experiment, we fit a model for the nouns with non-zero DLT score, encompassing 51,786 data points. The original experiment report having 45,038. In Table 2 we report the coefficients and significance level for all fixed effects of this model as well as the corresponding results of the original experiment. Like the original experiment, we find that the log(DLT) had a significant positive effect on reading time ($p < .01$). These two experiments demonstrate that parser bias did not skew the results substantially.

5 Conclusion

We introduced the Dundee Treebank—a new resource for corpus-based psycholinguistic experiments. The treebank is annotated in compliance with the Universal Dependencies scheme. We presented the design choices together with a batch of dependency parsing experiments.

We also partly replicated a study, which explores how a theory of sentence complexity, DLT, is reflected in reading times. We used manually assigned dependencies instead of parsed dependencies. Like the original experiment, we found both a small negative effect of DLT on all word and a significant positive effect of DLT on reading time for nouns with non-zero DLT score.

The treebank is made publicly available for research purposes.[4]

References

[1] Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Lindén, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Héctor Alonso Martínez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. Universal dependencies 1.1, 2015. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

[2] Maria Barrett and Anders Søgaard. Reading behavior predicts syntactic categories. In *CoNLL*, pages 345–249, 2015.

[3] Maria Barrett and Anders Søgaard. Using reading behavior to predict grammatical functions. In *CogACLL*, 2015.

[4] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[5] Bernd Bohnet. Top accuracy and fast dependency parsing is not a contradiction. In *COLING*, 2010.

[6] British National Corpus Consortium et al. British national corpus version 3, 2007.

[4] `https://bitbucket.org/lowlands/release`

[7] Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, pages 4585–4592, 2014.

[8] Vera Demberg and Frank Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210, 2008.

[9] Stefan L Frank. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *CogSci*, pages 1139–1144, 2009.

[10] Stefan L Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834, 2011.

[11] Edward Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126, 2000.

[12] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics, 2011.

[13] Alan Kennedy, Robin Hill, and Joël Pynte. The dundee corpus. In *ECEM*, 2003.

[14] Alan Kennedy and Joël Pynte. Parafoveal-on-foveal effects in normal reading. *Vision research*, 45(2):153–168, 2005.

[15] Scott A McDonald and Richard C Shillcock. Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, 43(16):1735–1751, 2003.

[16] Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. Syntactic and semantic factors in processing difficulty: An integrated measure. In *ACL*, pages 196–206, 2010.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[18] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[19] Nathaniel J Smith and Roger Levy. Fixation durations in first-pass reading reflect uncertainty about word identity. In *CogSci*, 2010.