Appendix for Modeling Input Uncertainty in Neural Network Dependency Parsing: Annotation Guidelines

1 Annotation Decisions

In this appendix we will give a short overview of annotation decisions. The Universal Dependencies English Web Treebank 2.1 (Silveira et al., 2014; Nivre et al., 2017) annotations are used as guidance for most annotation decisions. Since this treebank is sampled from different web domains, it already covers most phenomena occurring in Twitter data.

1.1 Tokenization

As a starting point, we used the tokenization from the previous annotation (Li and Liu, 2015). On top of this, we ran a simple rule-based tokenizer to make the data better suitable for syntactic annotation. Phrasal abbreviations (e.g. lol, smh) are treated as one token. We also included the normalization from the original corpora in the MISC column, which is manually corrected after tokenization (see Figure 1).

No sentence segmentation is performed on the input data because the Tweet-unit is inherent to this domain. Instead we use the `parataxis` relation to connect different utterances. The head of the first utterance is always the root, and all next utterance are dependents of this node, see Figure 2 for an example.

```
# text = damn im finna roll up again...
1    damn    Norm=damn
2    i       SpaceAfter=No;Norm=I
3    m       Norm=am
4    fin     SpaceAfter=No;Norm=going
5    na      Norm=to
6    roll    Norm=roll
7    up      Norm=up
8    again   Norm=again
9    ...     Norm=...
```

Figure 1: An example of tokenization in the CoNLL-U format (Only the 'ID', 'FORM' and 'MISC' column are shown here)

1.2 POS tags

Our parser does not make use of POS tags, but because they were already annotated and are closely related to the choice of dependency relations we corrected them during annotation. POS tags were first automatically mapped to universal tags (Petrov et al., 2012) and then manually corrected.

1.3 Unknown Words

If the annotator is unsure about the meaning of a word, other tweets containing the same word are searched and where necessary `www.urbandictionary.com` is consulted. If the annotator still could not understand the word, it is annotated as X with the `dep` relation. This only occurs five times in our data.

1.4 Emoticons, Emojis, URL's and Phrasal Abbreviations

Since words belonging to this category are often not syntactically connected, we annotate them as dependant of the head of the nearest utterance (see 1.1). The relations and POS tags used are similar to the English Web Treebank: emoticons and emojis are a SYMB connected with relation `discourse`, URL's are annotated as X with relation `appos` and phrasal abbreviations like 'lol' and 'smh' are considered to be an INTJ with the `discourse` relation.

1.5 Domain Specific Tokens

Mentions are used in Twitter to direct tweets towards a specific person/account. They consists of the '@' symbol followed by the targeted username. Because mentions are used to focus a Tweet to a specific user we annotated it as PROPN with the relation `vocative`.

Hashtags are used to specify the topic or mood of the Tweet. They are often located at the end of

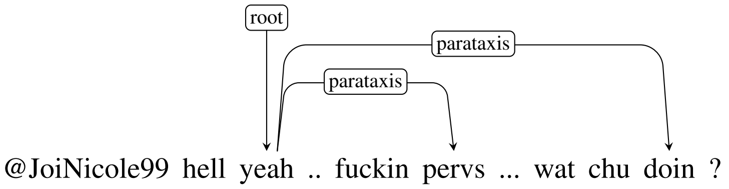

Figure 2: Annotation of the sentence "@JoiNicole99 hell yeah..fuckin pervs...watchu doin?", only relevant relations are shown.

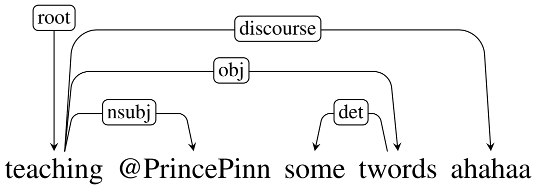

Figure 3: Annotation of the sentence "teaching @PrincePinn some twords ahahaa"

the Tweet. Their usage is similar to interjections in the English Web Treebank, so they are annotated accordingly as INTJ and discourse.

A retweet is indicated by the token 'RT', which is usually found at the beginning of the Tweet. We tag it with the X tag and the discourse relation.

Because these phenomena are often not syntactically connected to the sentence, we connect them to the root. Note that all of these phenomena can also be used in (syntactic) context, then they are annotated accordingly (see example in Figure3).

References

Chen Li and Yang Liu. 2015. Joint POS tagging and text normalization for informal text. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1263–1269.

Joakim Nivre, Zeljko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Silvie Cinková, Çağr Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, John Lee, Phuong Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Robert Ostling, Lilja Ovrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Simková, Kiril Simov, Aaron Smith, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van

Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jonathan North Washington, Mats Wirén, Tak-sum Wong, Zhuoran Yu, Zdeněk Zabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal Dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics ('UFAL), Faculty of Mathematics and Physics, Charles University.

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'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.