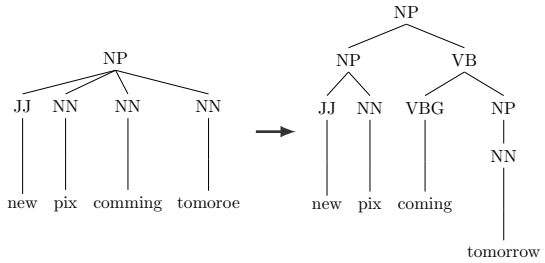
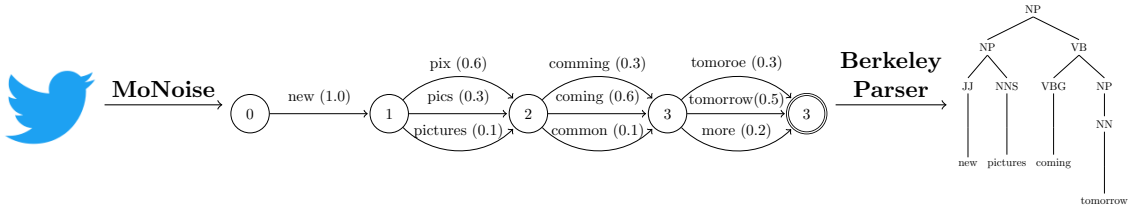


Corpus	Sents	Words/ sent	Unk%
WSJ (2-21)	39,832	23.9	4.4
EWT	16,520	15.3	3.7
Foster et al. (2011)	269	11.1	9.3
Li and Liu (2014)	2,577	15.7	14.1

Table 1: Some basic statistics for our training and development corpora. % of unknown words (Unk) calculated against the Aspell dictionary ignoring capitalization.



References

Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef Van Genabith. 2011. #hardtoparse: POS Tagging and parsing the Twitterverse. In *AAAI 2011 Workshop On Analyzing Microtext*. United States, pages 20–25.

Chen Li and Yang Liu. 2014. Improving text normalization via unsupervised model and discriminative reranking. In *Proceedings of the ACL 2014 Student Research Workshop*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 86–93. http://www.aclweb.org/anthology/P14-3012.

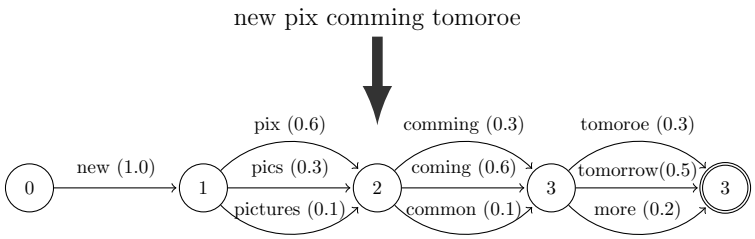


Figure 1: The output of the normalization model for the sentence “new pix comming tomoroe”.