- **Banana Split** [1] – splits compounds into only two parts.
- **jWordSplitter**[1]
- **Cdec compound-split** [2] – automatically converts to lowercase.

Tool	**Vocabulary size** (lowercase)	**Reduction**
baseline	188 812 (182 837)	–
Banana Split	145 233 (133 077)	23 % (27 %)
jWordSplitter	124 514 (112 544)	35 % (38 %)
Cdec compound-split	100 888	47 %
Cdec compound-split + truecasing	101 278	46 %

Table 1: Vocabulary size and its reduction after splitting

Tool	**Testset**		**Devset**	
	BLEU	**METEOR**	**BLEU**	**METEOR**
baseline	22.38	25.31	22.81	25.50
Banana Split	22.26	25.34	22.96	25.74
jWordSplitter	**22.49**	25.55	**23.02**	25.83
Cdec compound-split	22.31	25.46	22.93	25.84
Cdec compound-split + truecasing	22.32	**25.57**	23.01	**25.91**

Table 2: Translation from German to Czech (561k sent. train, 3k dev, 3k test)

[1] `https://github.com/danielnaber/jwordsplitter`

Bibliography

[1] Niels O. *Evaluation of the BananaSplit Compound Splitter*. Technical report, Seminar für Sprachwissenschaft, Eberhard-Karls-Universität Tübingen, 2006.

[2] Dyer, C., Weese, J., Setiawan, H., Lopez, A., Ture, F., Eidelman, V., Ganitkevitch, J., Blunsom, P. a Resnik, P. Cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations* (Stroudsburg, PA, USA, 2010), ACLDemos '10, Association for Computational Linguistics, pp. 7–12.