Rudolf Rosa, Martin Popel, Ondřej Bojar,
David Mareček, Ondřej Dušek
{**rosa**,popel,bojar,marecek,odusek}@ufal.mff.cuni.cz

Moses & Treex Hybrid MT Systems Bestiary

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

DMTW, Lisbon, 21 October 2016

English-to-Czech MT

- Individual Systems
  - Moses & Factored Moses
  - Treex & TectoMT
- 8 System Combinations
  - description
  - evaluation
- Conclusion

WARNING

- an overview paper
  - we review already existing work
- described systems come from other authors
  - still, usually an overlap with authors of this paper
- BLEU scores
  - numbers taken from previously published papers
  - various datasets, various approaches…
  - gain in BLEU vs vanilla Moses/TectoMT baseline
- all references in the paper

Moses

- well-known phrase-based statistical MT system
- extremely poor in explicit linguistic knowledge
  - tokenization, truecasing… (and not much more)
- Czech is a hard target language
  - morphologically rich
    - unigram sparsity
  - free word order, long distance dependencies
    - n-gram sparsity

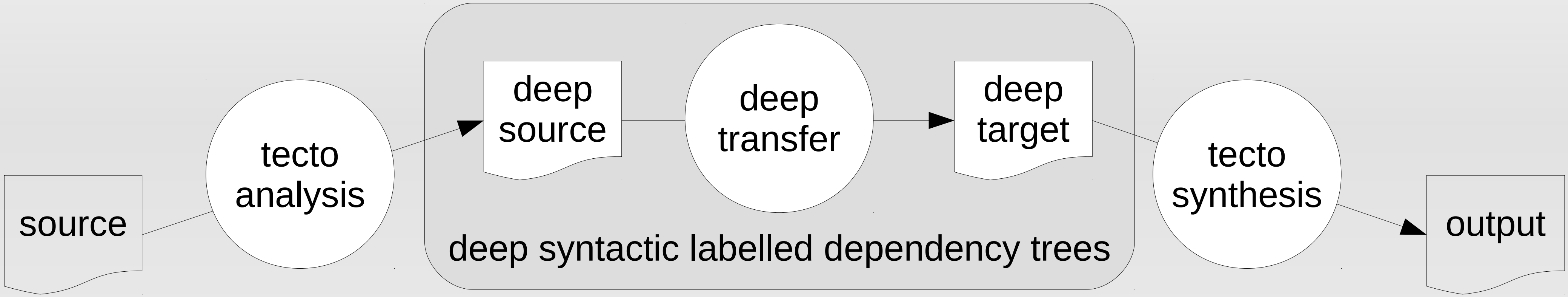
Factored Moses

- factored representation of Czech
  - word = (word form, part of speech tag)
- additional language model on PoS tags
  - helps overcome data sparsity
- PoS tags provided by Treex

Treex

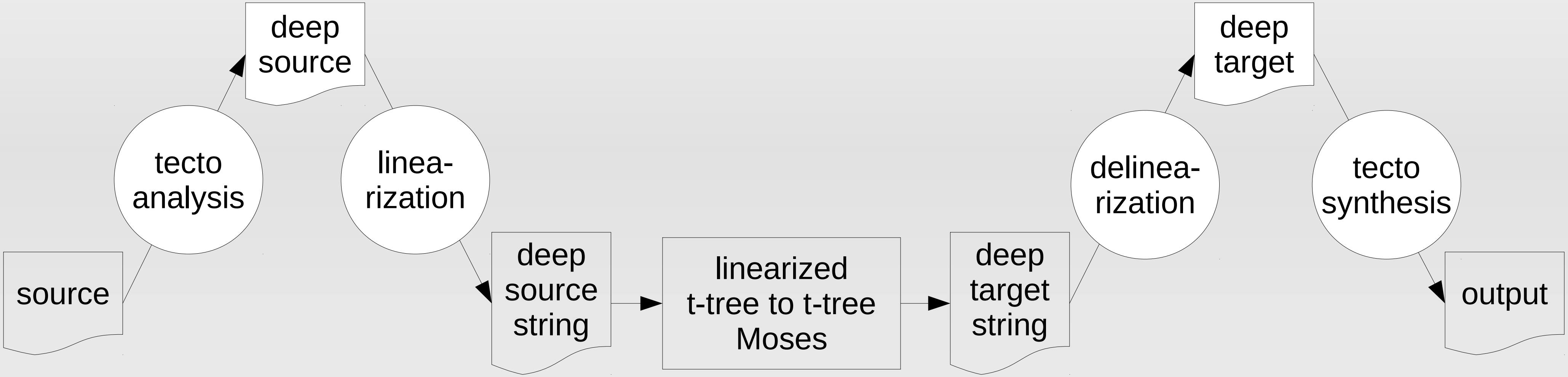
- linguistically motivated modular NLP framework
  - individual components (blocks) – tagger, parser…
    - both rule-based and machine-learned
  - flexibly combined into complex processing pipelines
- several layers of linguistic abstraction
  - based on Prague Dependency Treebank
  - a-tree (surface syntax), t-tree (deep syntax)…
  - rich linguistic annotation for each token
  - abstraction over inflection, word order, aux words…
- multilingual

TectoMT



- hybrid MT system implemented in Treex
- deep transfer of t-trees
  - t-tree structure and grammatemes copied 1:1
    - isomorphic transfer
  - lemmas and formemes generated by ML models
- typically worse than Moses by ~4 BLEU points

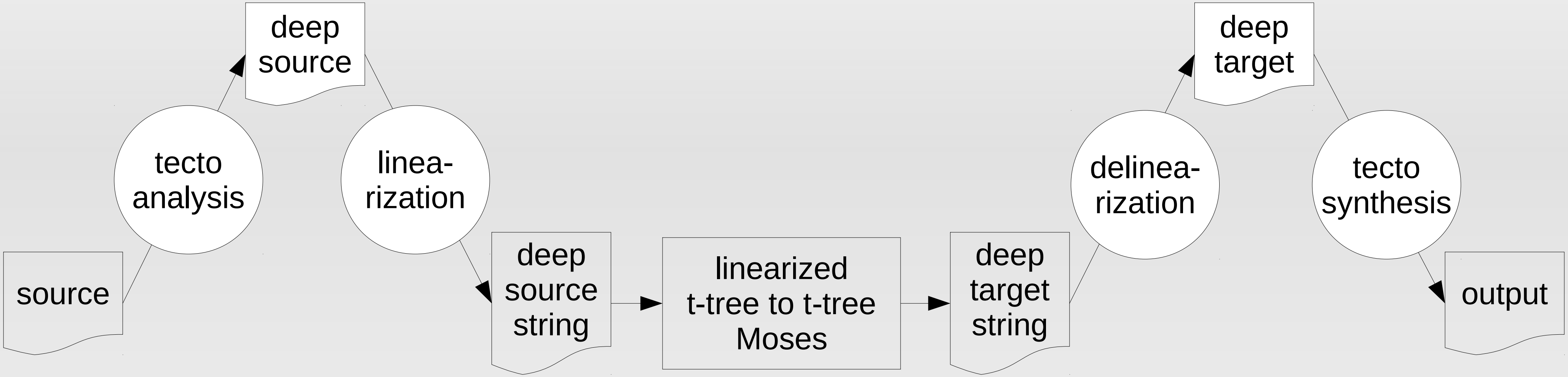
TectoMoses



1. source t-tree → string of lemmas and formemes
2. string translated by specially trained Moses
3. target string → target t-tree
   - dependencies projected through alignment

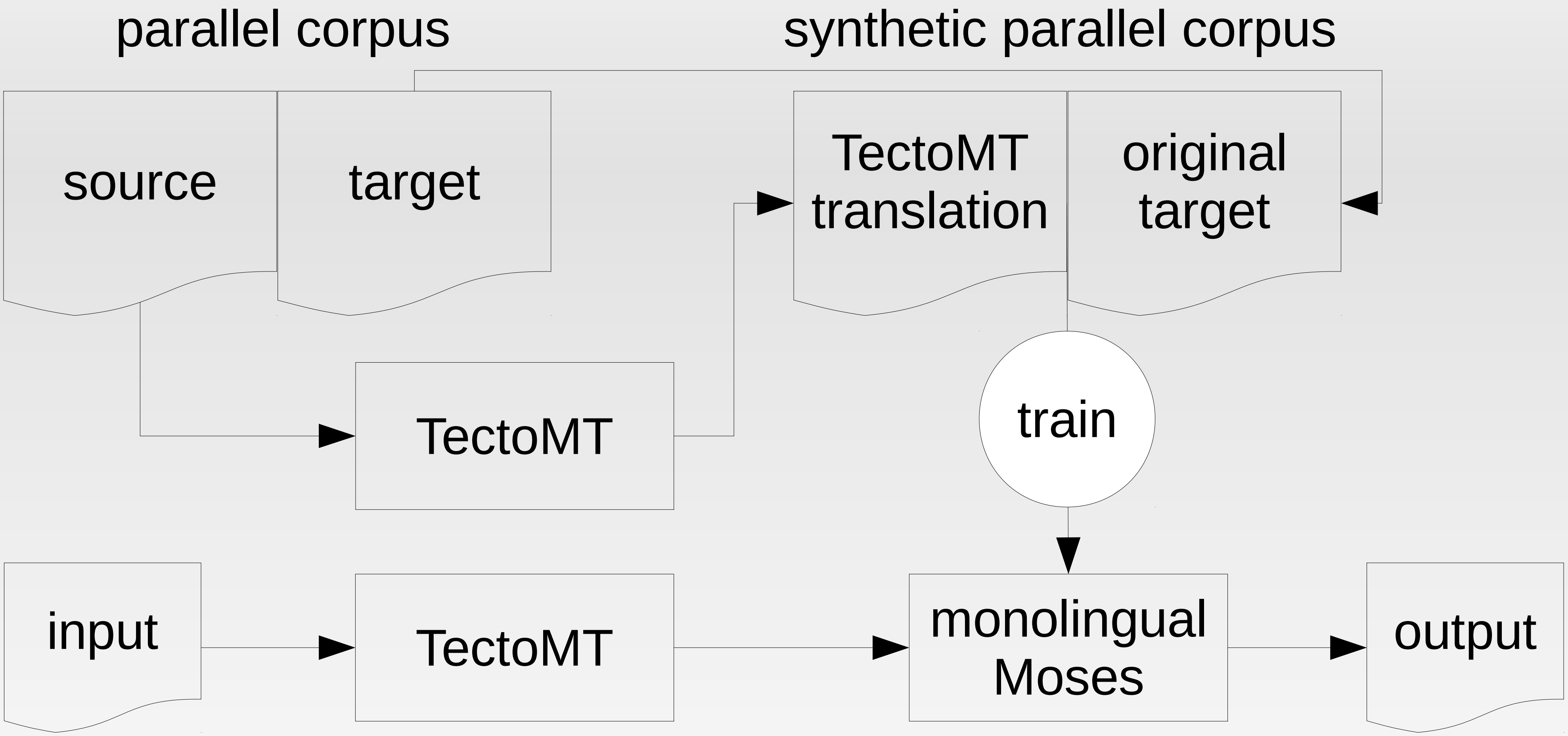
- supports non-isomorphic transfer

TectoMoses



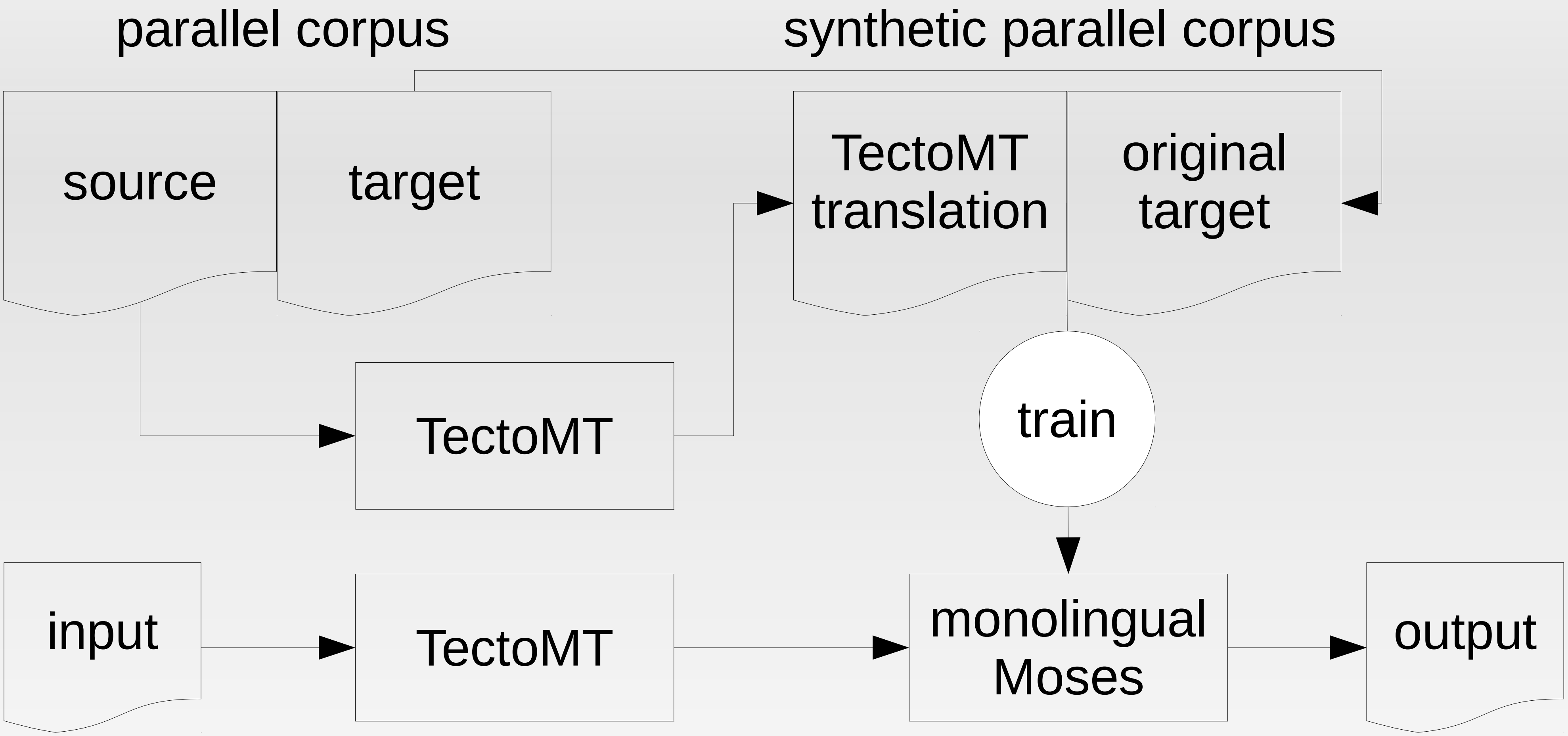
- -2.2 BLEU vs vanilla TectoMT baseline

PhraseFix: Moses post-editing



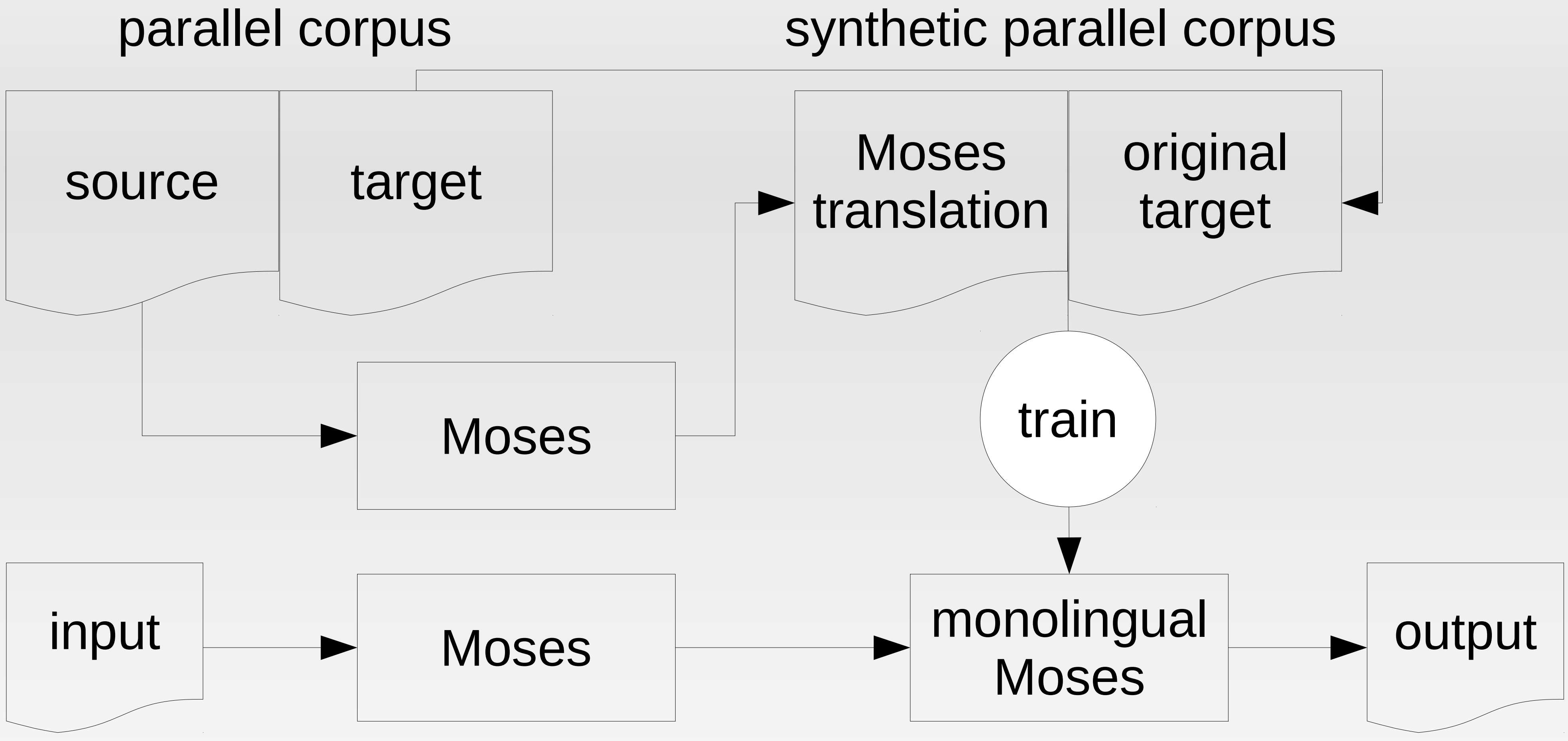
- train Moses to post-edit outputs of TectoMT

PhraseFix: Moses post-editing



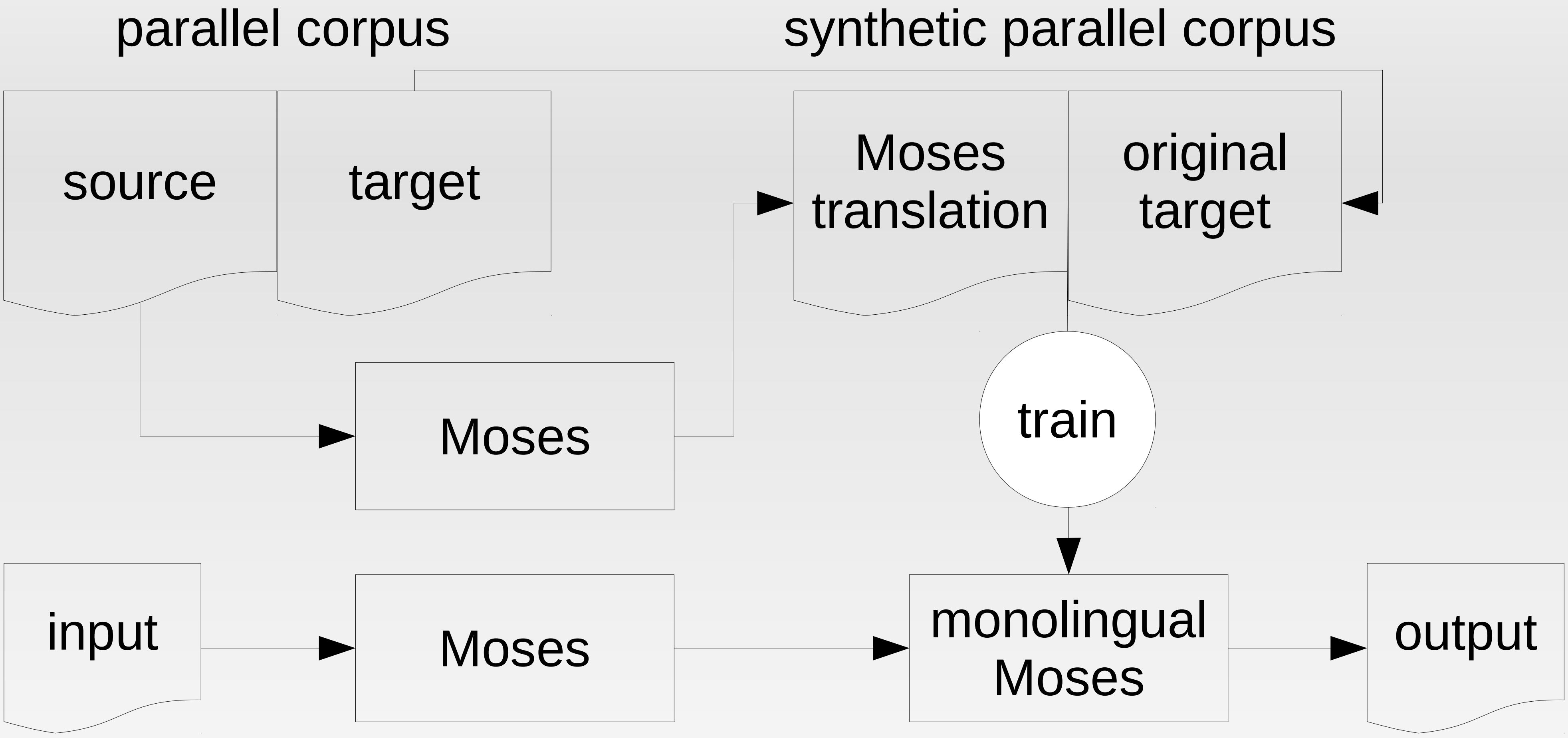
- up to +3.2 BLEU vs TectoMT
  - but still worse than Moses

Moses + Moses post-editing



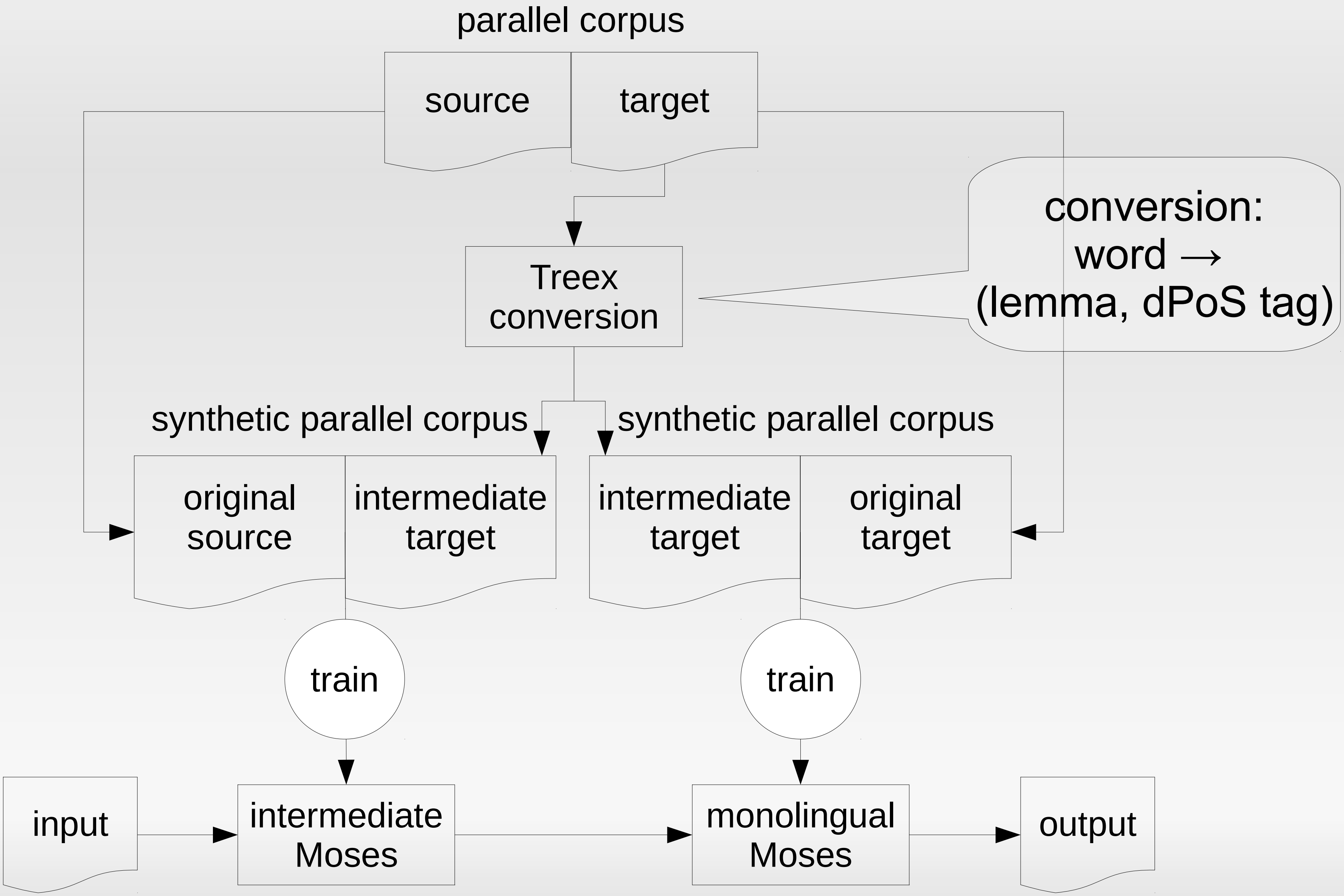
- train Moses to post-edit outputs of Moses

Moses + Moses post-editing

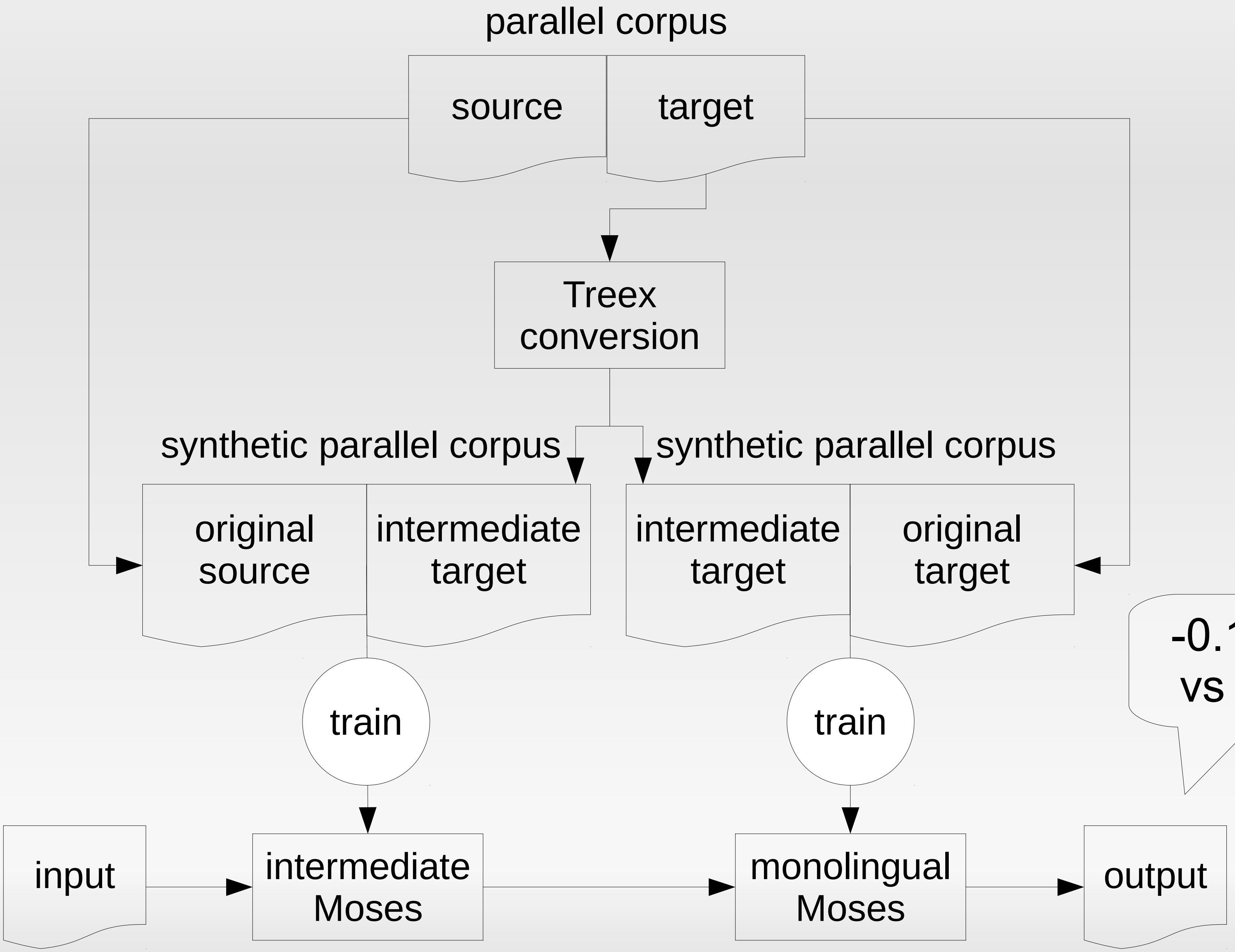


- -0.1 BLEU vs Moses

TwoStep Moses

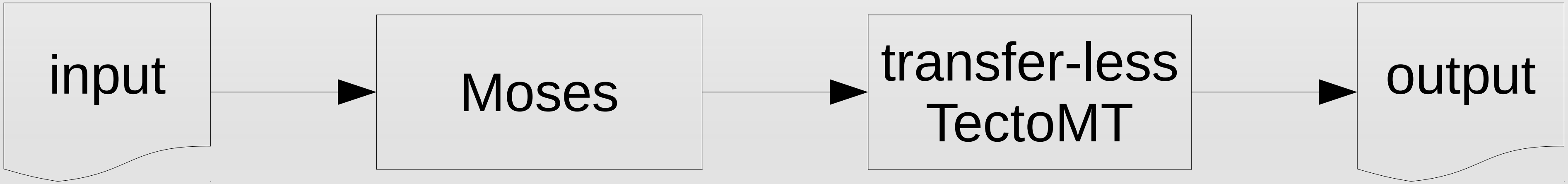


TwoStep Moses



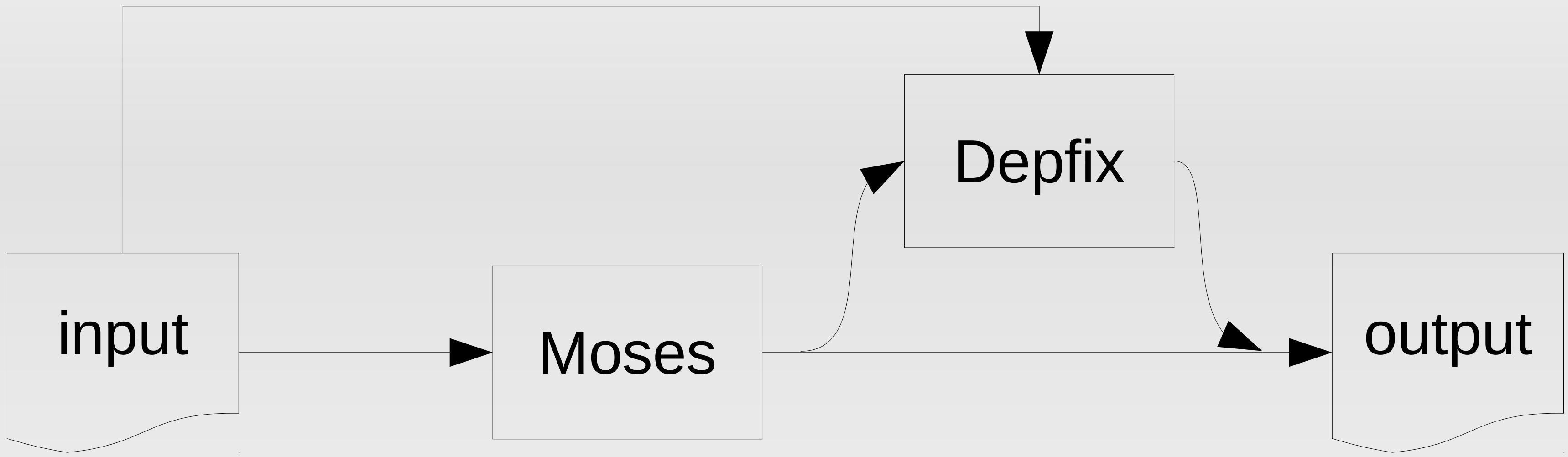
-0.1 BLEU vs Moses

Moses + TectoMT post-editing



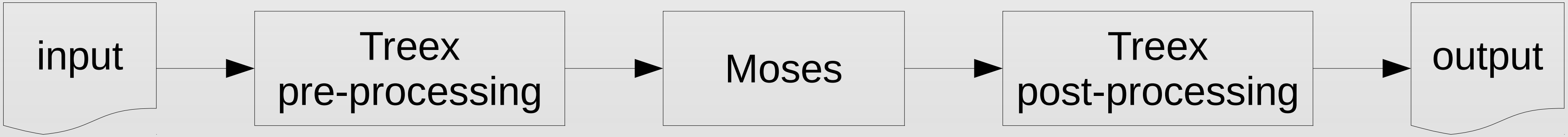
- transfer-less TectoMT (monolingual)
  1. t-analysis (text → t-tree)
  2. t-synthesis (t-tree → text)
- +2.4 BLEU vs TectoMT, -2.4 BLEU vs Moses
  - occasionally fixes some grammatical agreement etc.
  - t-analysis very noisy → t-synthesis also noisy

Moses + Depfix post-editing



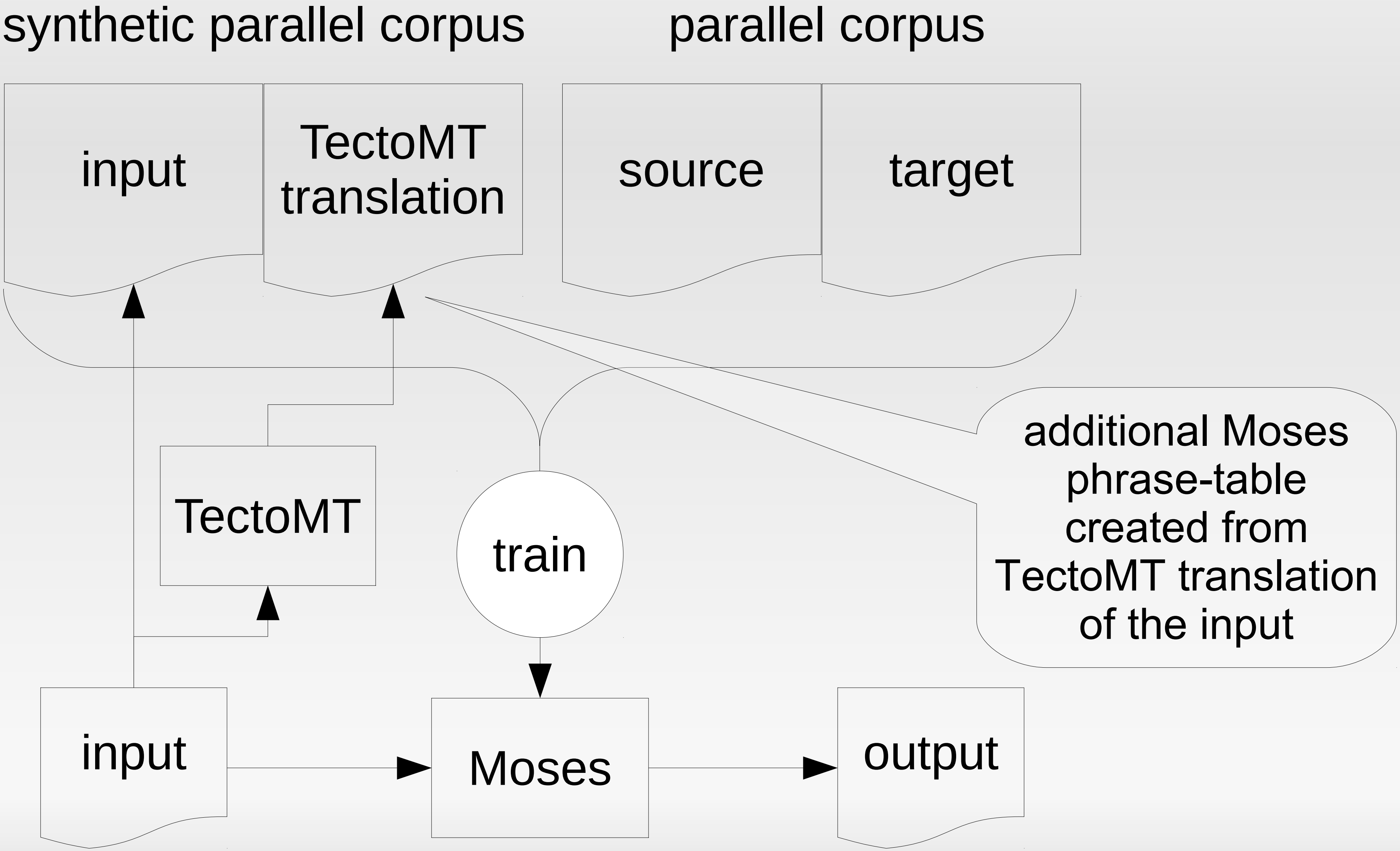
- Depfix = dozens of rule-based Treex blocks
  - tries to fix only sure errors, does not touch the rest
    - morphological agreement, lost negation…
  - also analyzes the input for additional information
- up to +0.4 BLEU vs Moses (usually around +0.1)

Moses + Treex pre-/post-processing



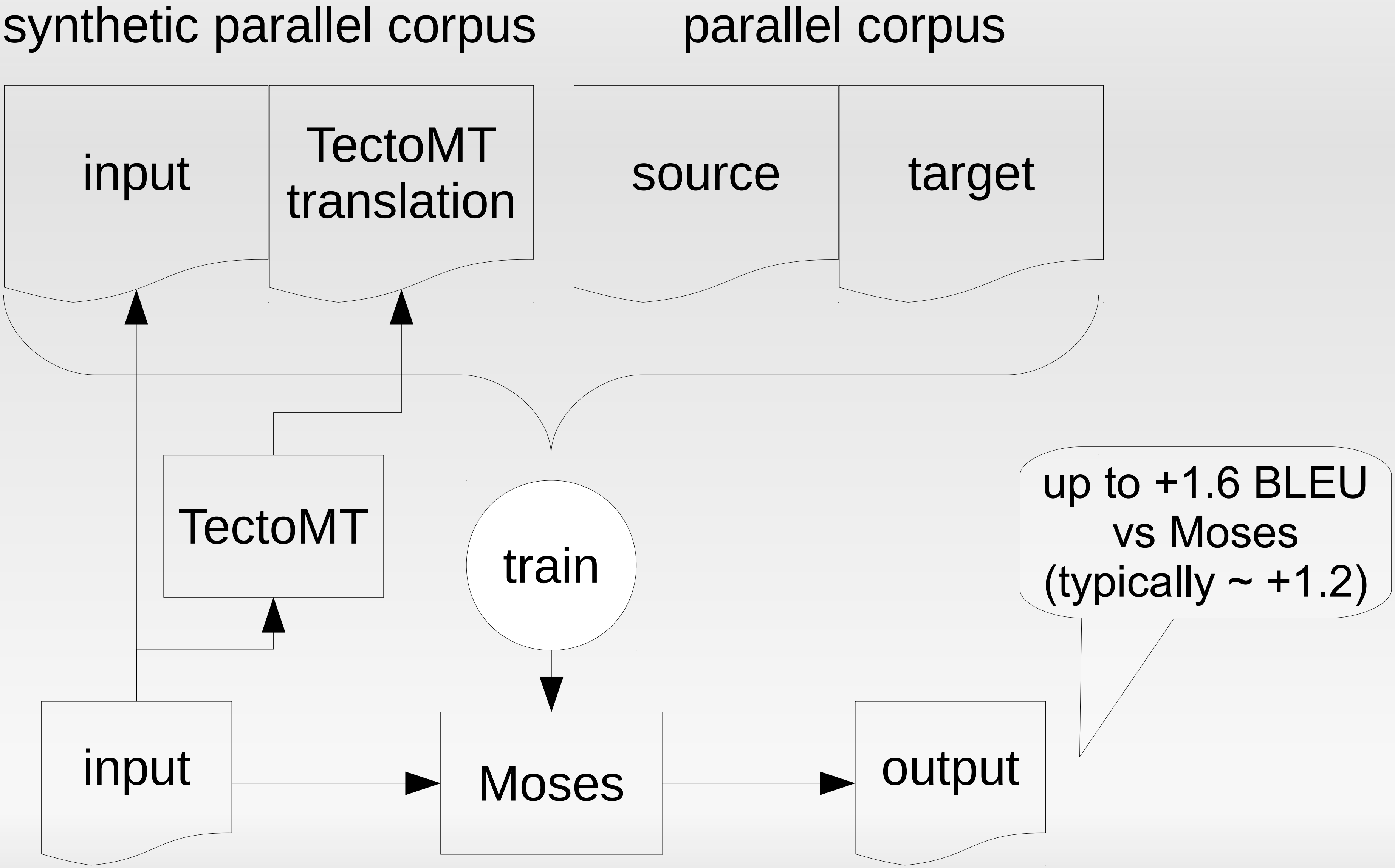
- handle phenomena hard for Moses
  - remove articles, mark subjects, reorder…
  - force-translate named entities using a gazetteer
  - preserve URLs, e-mail addresses, filenames…
- +0.4 BLEU vs Moses

Two-headed Chimera



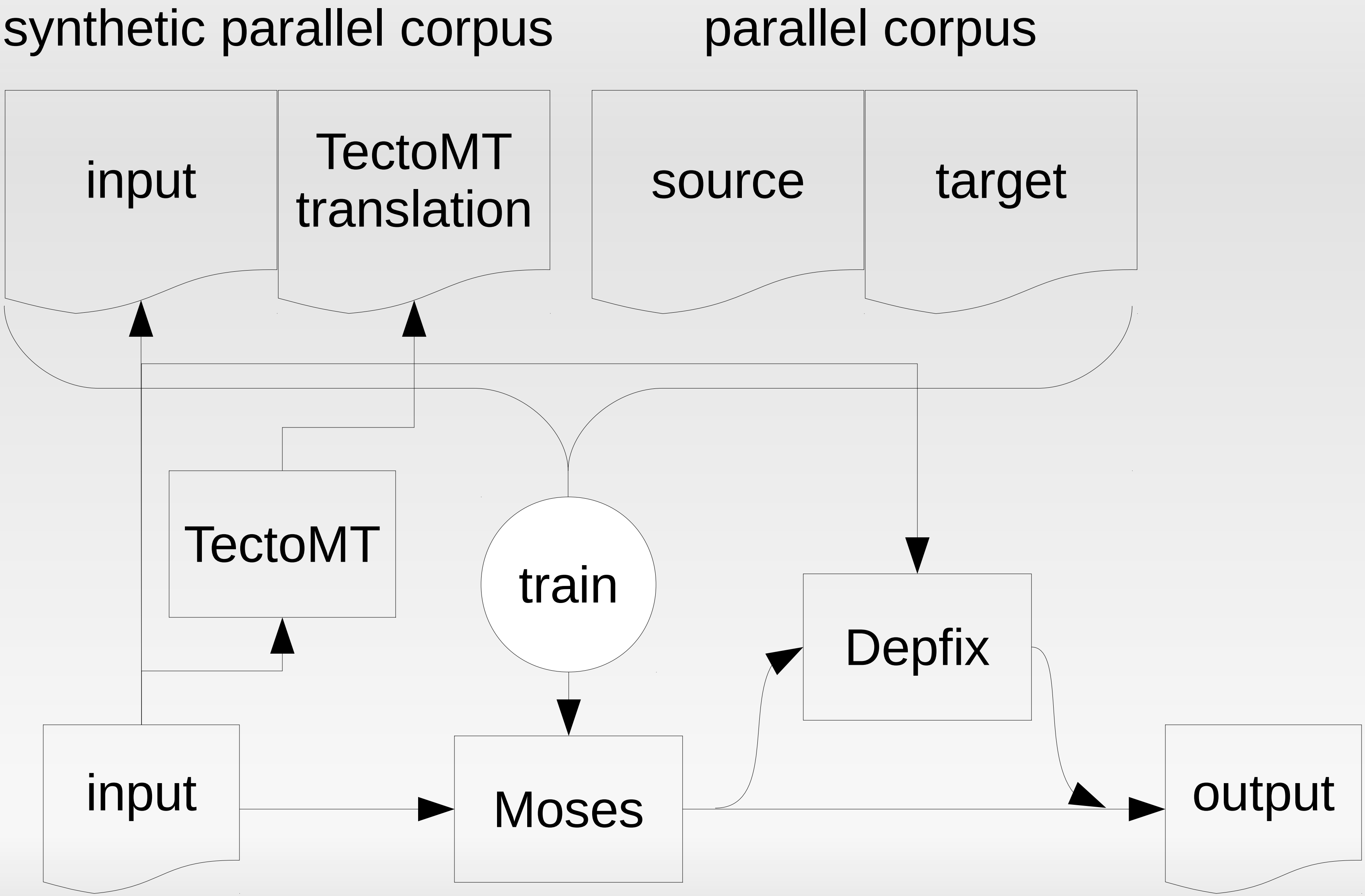
synthetic parallel corpus
parallel corpus
input
TectoMT translation
source
target
TectoMT
train
additional Moses phrase-table created from TectoMT translation of the input
input
Moses
output

Two-headed Chimera

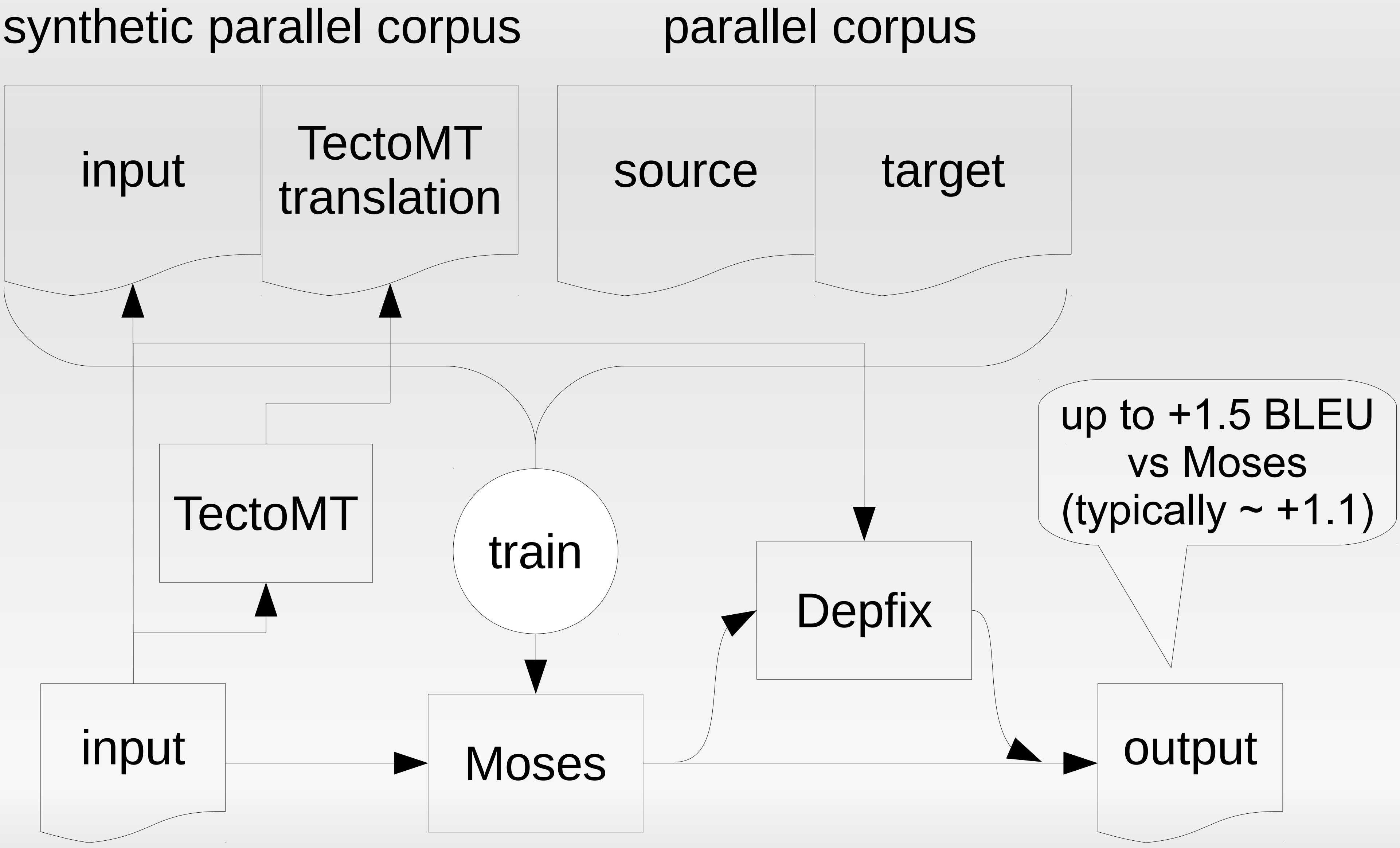


up to +1.6 BLEU vs Moses (typically ~ +1.2)

Three-headed Chimera



Three-headed Chimera



Conclusion

- range of Treex/TectoMT & Moses combinations
- some brought significant improvements
  - +1.5 BLEU Chimera (two-headed/three-headed)
    - best system in WMT 2013, 2014, 2015
  - +0.5 BLEU Treex pre-/post-processing
  - +0.4 BLEU Depfix
- other currently not very useful
  - future potential?

Thank you for your attention

Rudolf Rosa, Martin Popel, Ondřej Bojar,
David Mareček, Ondřej Dušek

{rosa,popel,bojar,marecek,odusek}@ufal.mff.cuni.cz

Moses & Treex Hybrid MT Systems Bestiary

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

http://ufal.mff.cuni.cz/rudolf-rosa/