

Machine Translation between Languages with Significant Word Reordering and Rich Target-side Morphology

20th Week of Doctoral Students, June 3rd, 2011

ÚFAL, Charles University in Prague

Bushra Jawaid

RNDr. Ondřej Bojar (PhD. Advisor)



Language Pair & Properties

- Language Pair → English-Urdu
- English is SVO language and has strict word order.
- Urdu is restricted **free** word order language and mostly follows SOV structure by default.

English Sentence:	I understand English and Urdu?					
Urdu Translation:	ہوں	سمجھتی	اُردُو	اور	انگریزی	میں
Transliteration:	meñ	angrezī	aor	Urdū	samjhte	hūñ
Gloss:	I	English	and	Urdu	understand	(Auxiliary)

Language Pair & Prop (Cont)

- Urdu has concatenative inflective morphological system.
- For example, verbs in Urdu inflects for tense, mood, aspect, gender and number.
- Table below shows three different masculine forms of verb (be made)

	Root	Infinitive	Oblique
Intransitive/ (di) Transitive	bən بن	bəna بنا	bəne بنے
Direct Causative	bəna بنا	bəna بنا	bəne بنے
Indirect Causative	bəna بنا	bəna بنا	bəne بنے

Research Focus ..

- Exploring methods and techniques when translating into the direction of morphologically richer languages.
- Reduce the word order differences in source and target languages.
- Main motivation:
 - Model the problem of reordering.
 - Deal with word form choice separately.
 - Improve generalization.

Possible Solutions

Translate+Generate (T+T+G) Setup (Bojar et al., 2010):

English	Czech	
Form	Form	← +LM
Lemma	Lemma	+LM
Morphology	Morphology	+LM

Issues with this setup:

- Factors in Moses synchronous → all factors have to be fully constructed before main search.
- Many possible options of lemma, tag and final word form → Pruning strikes hard.

Possible Solutions (Cont) ..

- Translation options of German word “haus”, (Koehn et al. 2007)

- Translation: Mapping lemmas

{ ?|house|?|?, ?|home|?|?, ?|building|?|?, ?|shell|?|? }

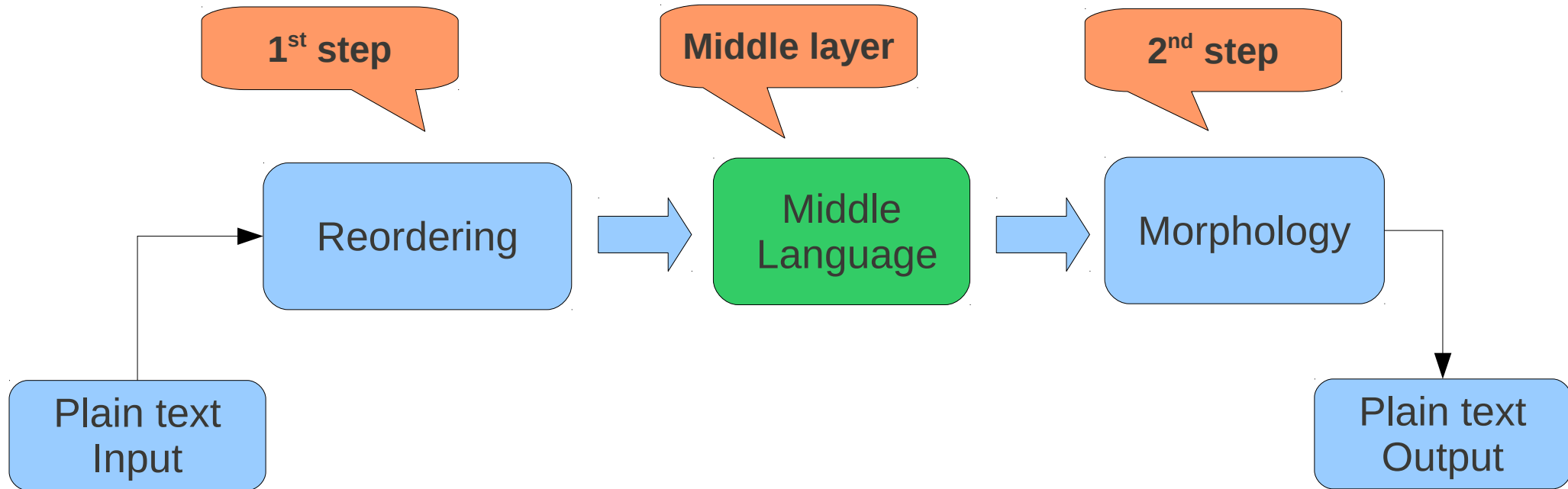
- Translation: Mapping morphology

{ ?|house|NN|plural, ?|home|NN|plural, ?|building|NN|plural, ?|shell|NN|plural, ?|house|NN|singular,... }

- Generation: Generating surface forms

{ houses|house|NN|plural, homes|home|NN|plural, buildings|building|NN|plural, shells|shell|NN|plural, house|house|NN|singular, ... }

Two-Step Architecture..



(Fraser, 2009) and (Bojar, 2010)

Possible Solutions (Cont) ..

- **Two-Step Setup** (to avoid explosion of translation options):
- First step translates from source to augmented lemmatized target word.
- Monolingual features are **not** represented, for example the gender for adjectives.

Src	good	book
Mid	A1XX.اچھا	NSNX.کتاب
Gloss	adj+1stdeg...good	noun+sg+nom...book

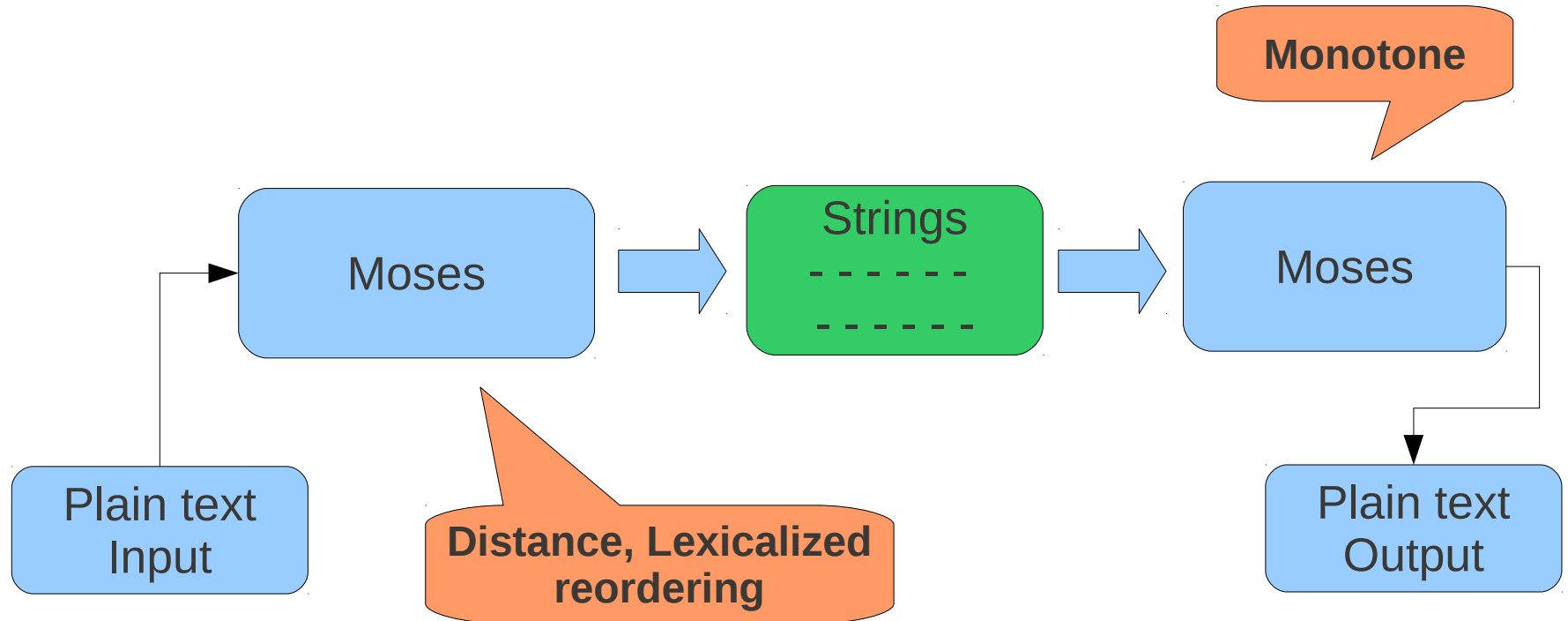
Possible Solutions (Cont) ..

- The second step is monotone translation from lemmatized target word to fully inflected target word.

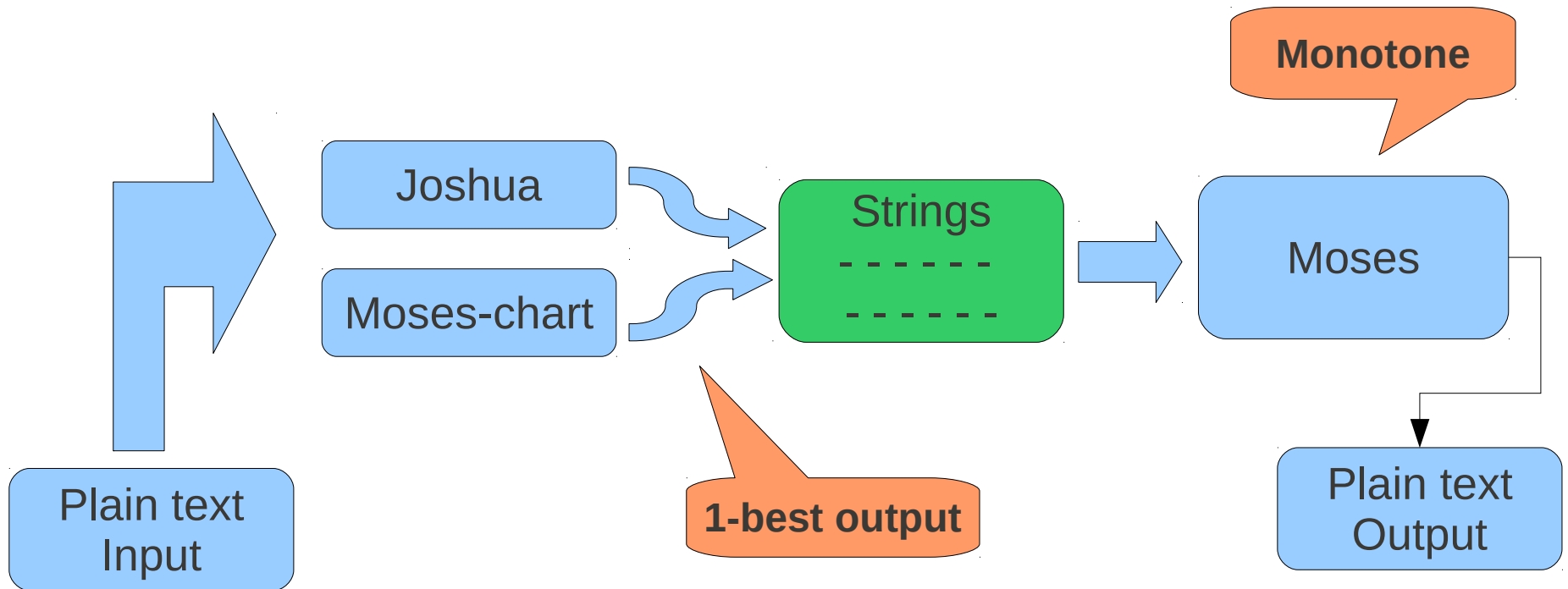
Src	good	book
Mid	A1XX.اچھا	NSNX.کتاب
Gloss	adj+1stdeg...good	noun+sg+nom...book
Out	اچھی(achi)	کتاب (kitab)

Idea behind 2-step architecture → Model target-side morphology separately if not dependent on source morphology.

Basic Two-Step Setup..



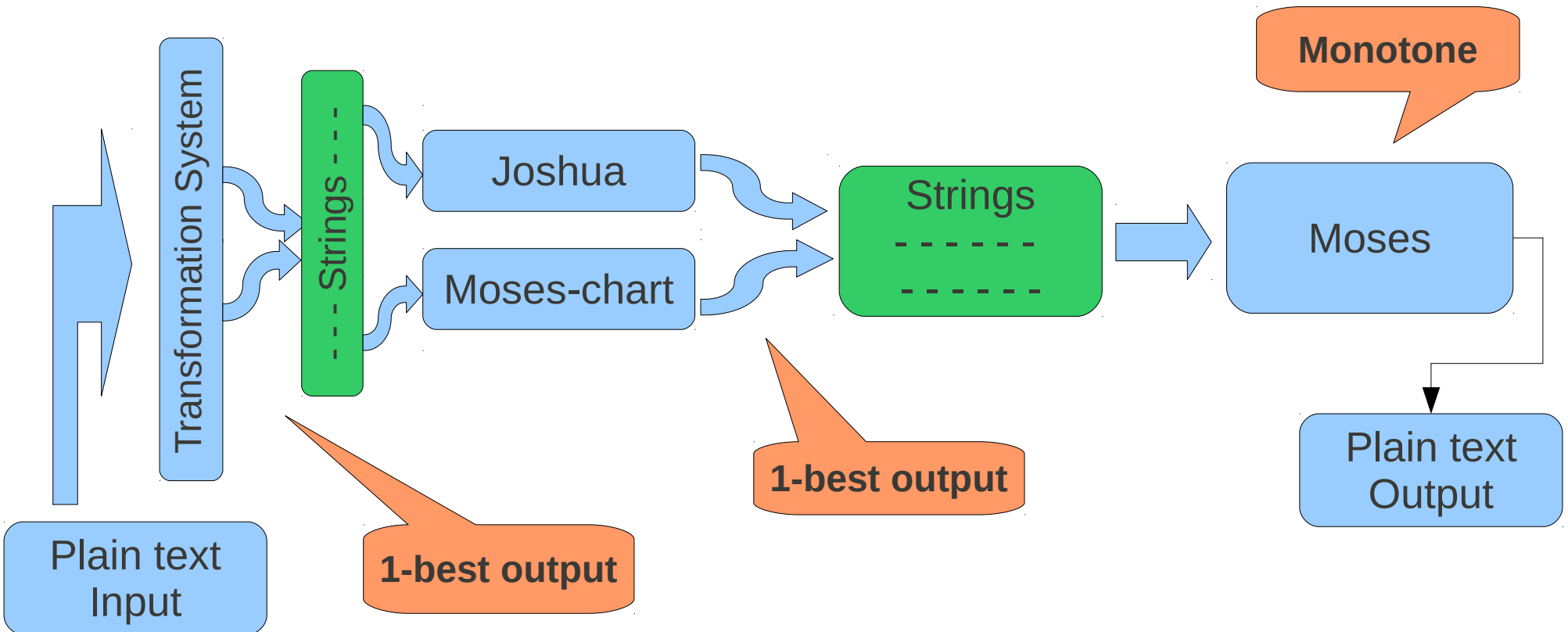
Two-Step Variants



1. Reordering options:

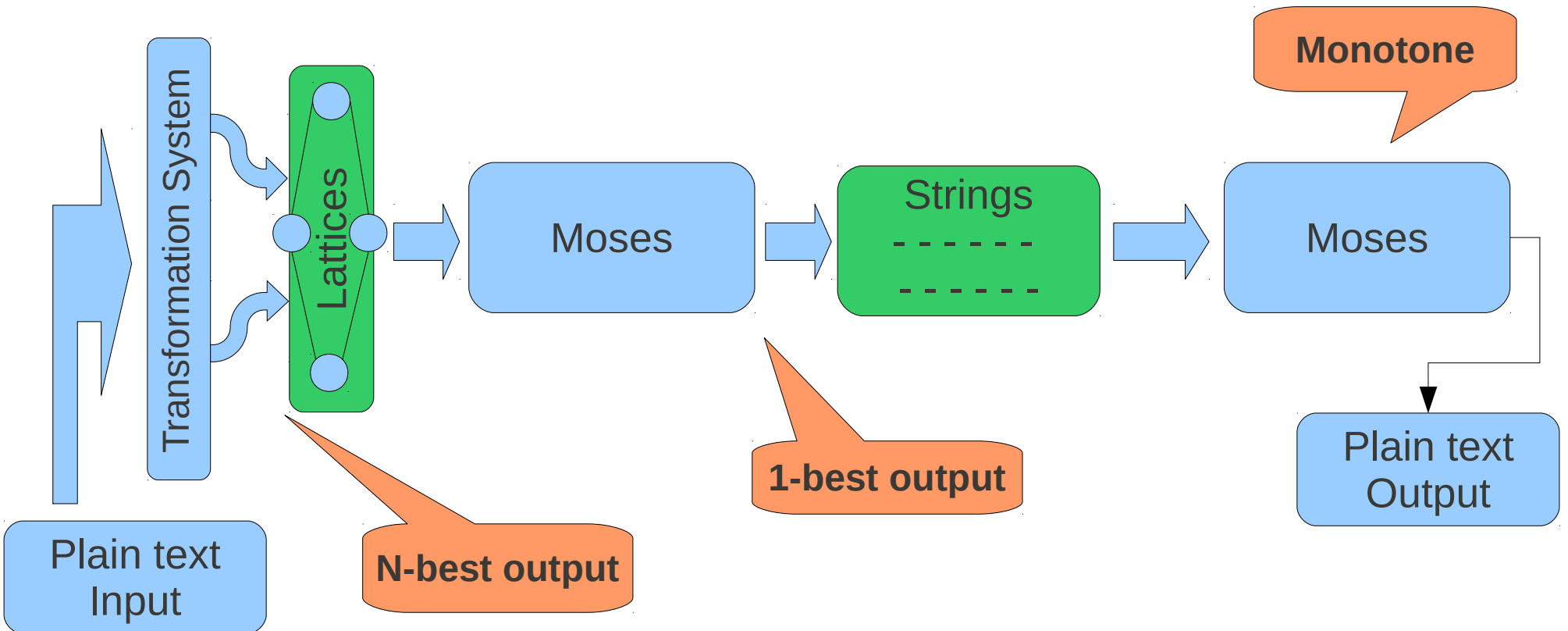
- Using moses-chart or Joshua or manual reordering on 1st step for improved reordering.
- Moses-chart and joshua are hierarchical, i.e. allow block movements.

Two-Step Variants (Cont ..)



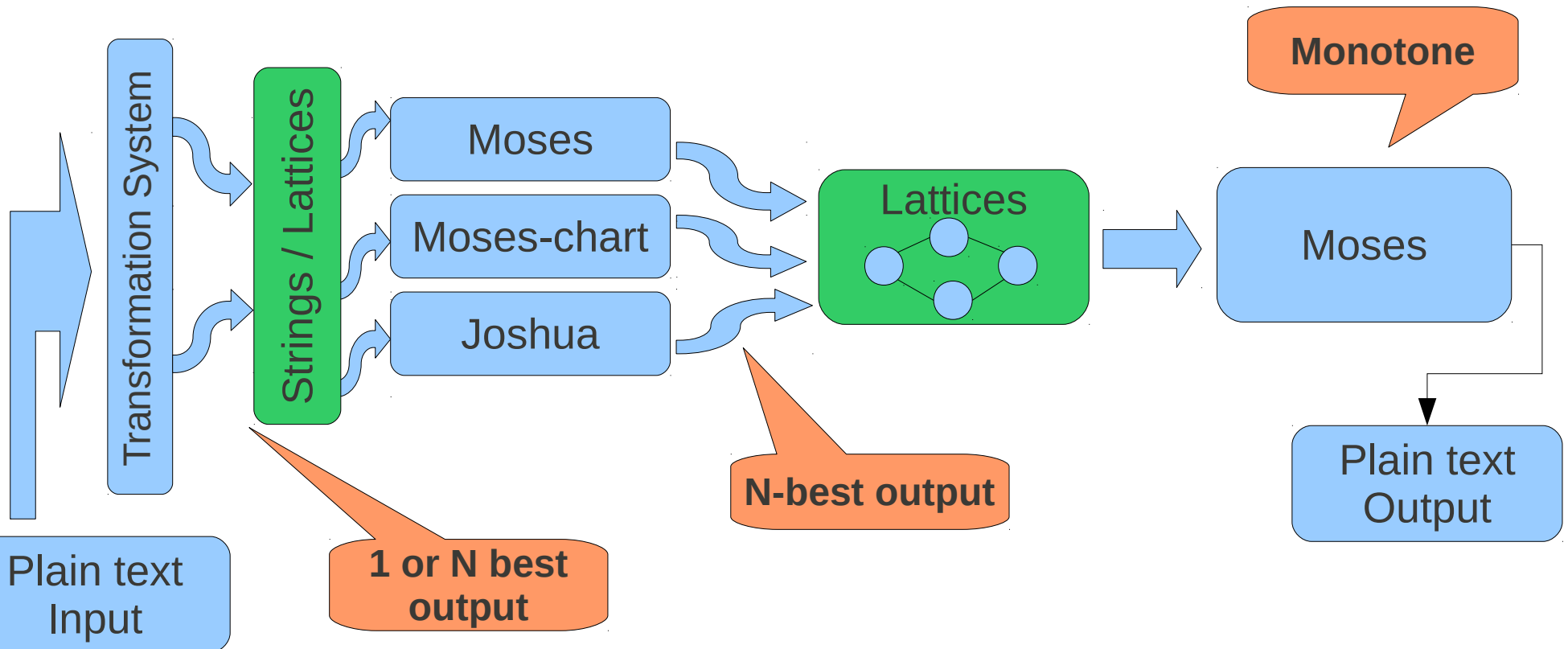
- Pre-reorder input sentences using Transformation system (Jawaid, 2010) and pass 1-best reordered output to 1st layer.

Two-Step Variants (Cont ..)



- Generate input lattice from multiple reorderings of each sentence.
- Use of lattices (Niehues et al. 2009) and (Bisazza et al. 2010).

Two-Step Variants (Cont ..)

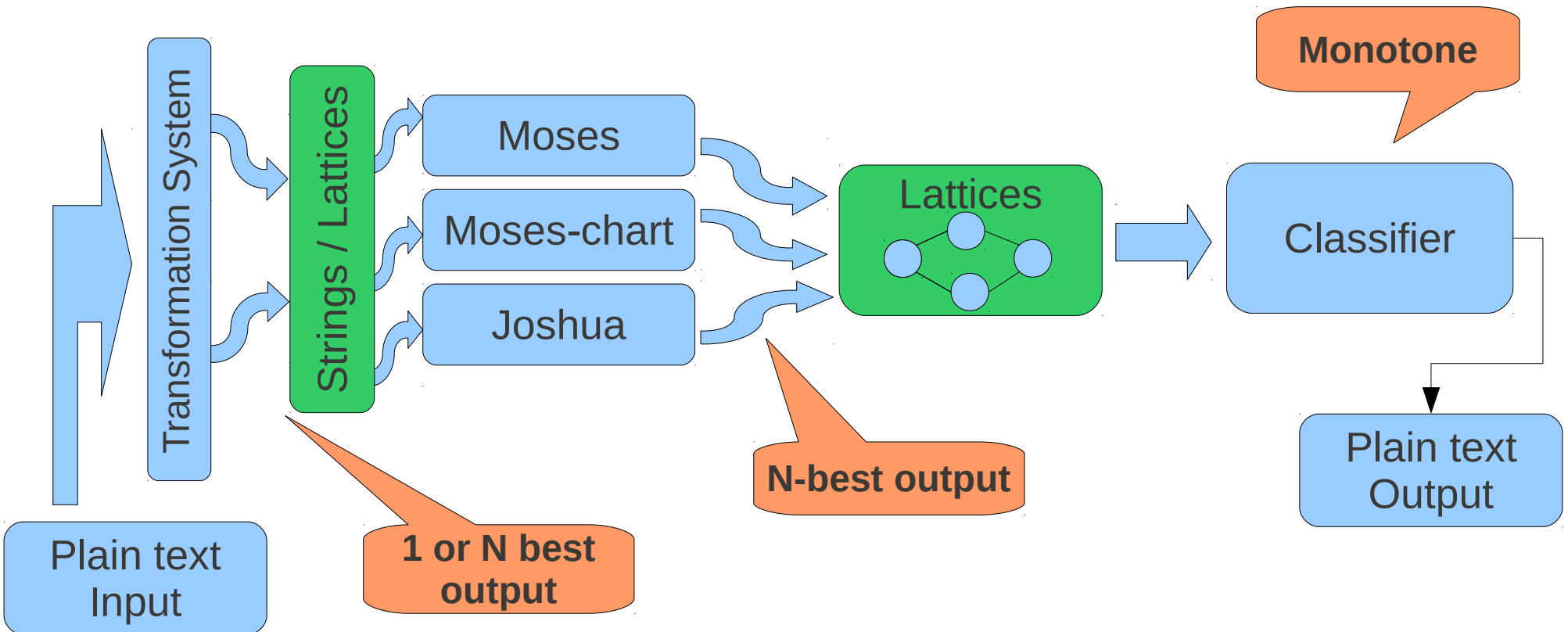


2. Middle Layer options:

Passing lattices of possible hypothesis from 1st step to 2nd step instead of passing hypothesis of simple string.

Multiple reorderings are considered and 2nd step is free to choose the one that is the easiest to inflect.

Two-Step Variants (Cont ..)



3. 2nd Layer options:

Adding a classifier on 2nd step to get the best hypothesis.

Main Issues ..

- Urdu is under-resourced language.
- Current research work:
 - Finding and Improving Taggers
 - Collecting tools such as tagger and morphological analyzer for Urdu.
 - Trying to combine the taggers to improve precision.
 - Need to merge the different tagsets.
 - Collecting more data.

Questions?

Feel free to ask questions.

