共生社会特論

第6回　派生文法とウイグル語機械翻訳

2016年1月24日

内容

- 日本語形態素解析
  - 派生文法
  - 派生文法に基づく形態素解析システム
- 日本語ーウイグル語機械翻訳
  - 日本語とウイグル語の共通点
  - 派生文法に基づく翻訳
  - 実験

日本語形態素解析

- 日本語処理において不可欠
- 様々なシステム

システム	文法
JUMAN [長尾ら ’93]	益岡・田窪 [’92]
茶筅 [松本ら ’96]	学校文法 [橋本’48]
MeCab [工藤ら ’04]	

IPA 品詞体系

動詞の活用形の処理

活用処理

学校文法による解析

書か	力行五段活用 動詞未然形
せ	助動詞未然形
られ	助動詞連用形
ます	助動詞終止形

- 動詞の細かな分類が必要
- 活用形のチェックが必要

活用処理手法

- 活用形展開方式
- 活用語尾分離方式
- 活用語尾展開方式

[久光ら ’92]

活用処理

活用形展開方式

活用変化した形を
すべて辞書へ登録

書か ＋ れる
書き ＋ ます
書く ＋ とき
書け ＋ ば
書こ ＋ う
書い ＋ た

活用処理手法

- 活用形展開方式
- 活用語尾分離方式
- 活用語尾展開方式

[久光ら '92]

活用処理　　活用語尾分離方式

活用語尾を分離して辞書へ登録

書＋か＋れ＋る
書＋き＋ま＋す
書＋く＋とき
書＋け＋ば
書＋こ＋う
書＋い＋た

活用処理手法

- 活用形展開方式
- 活用語尾分離方式
- 活用語尾展開方式

[久光ら '92]

活用処理　　活用語尾展開方式

語幹末尾の子音を
語幹の先頭に付加

書 ＋ かれ ＋ る
書 ＋ きま ＋ す
書 ＋ く ＋ とき
書 ＋ けば
書 ＋ こう
書 ＋ いた

活用処理手法

- 活用形展開方式
- 活用語尾分離方式
- 活用語尾展開方式
  [久光ら ’92]

活用処理

音韻論的手法

文字単位ではなく
音韻単位で解析

書k ＋ are ＋ ru
書k ＋ imas ＋ u
書k ＋ u ＋ toki
書k ＋ eba
書k ＋ ou
書 ＋ ita

活用処理手法
- 活用形展開方式
- 活用語尾分離方式
- 活用語尾展開方式

[久光ら ’92]

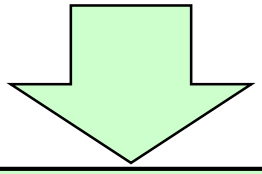
音韻論的手法
- Bloch [’46]
- 寺村 [’84]

音便形に対して特別な処理が必要

派生文法　[清瀬 '89]

日本語は活用しない

- 音韻論的文法
- 日本語の膠着語的性質に着目
- 動詞の語形変化 ＝ 語幹＋接尾辞



単純かつ体系的な文法

日本語処理に適しているのでは？

派生文法に基づく解釈

五段活用動詞	一段活用動詞
書かれる	食べられる
書きます	食べます
書く	食べる

派生文法に基づく解釈

五段活用動詞	一段活用動詞
書kareru	食berareru
書kimasu	食bemasu
書ku	食beru

派生文法に基づく解釈

五段活用動詞	一段活用動詞
書k-(r)are-(r)u	食be-(r)are-(r)u
書k-(i)mas-(r)u	食be-(i)mas-(r)u
書k-(r)u	食be-(r)u

派生文法に基づく解釈

子音幹動詞	母音幹動詞
書k-(r)are-(r)u	食be-(r)are-(r)u
書k-(i)mas-(r)u	食be-(i)mas-(r)u
書k-(r)u	食be-(r)u

派生接尾辞　新たな語幹を派生する接尾辞

統語接尾辞　動詞形を形成する接尾辞

派生文法に基づく解釈

子音幹動詞	母音幹動詞
書k-(r)are-(r)u	食be-(r)are-(r)u
書k-(i)mas-(r)u	食be-(i)mas-(r)u
書k-(r)u	食be-(r)u

派生接尾辞　新たな語幹を派生する接尾辞

統語接尾辞　動詞形を形成する接尾辞

接続規則

1. 子音幹に接続すると、連結子音が欠落する
2. 母音幹に接続すると、連結母音が欠落する

派生文法に基づく解釈

子音幹動詞	母音幹動詞
書k-(r)are-(r)u	食be-(r)are-(r)u
書k-(i)mas-(r)u	食be-(i)mas-(r)u
書k-(r)u	食be-(r)u

派生接尾辞　新たな語幹を派生する接尾辞

統語接尾辞　動詞形を形成する接尾辞

接続規則

1. 子音幹に接続すると、連結子音が欠落する
2. 母音幹に接続すると、連結母音が欠落する

派生接尾辞・統語接尾辞

	役割	接尾辞	用例	
			子音幹	母音幹
派生接尾辞	使役	-(s)ase-	kak-ase-	tabe-sase-
	受身・尊敬・可能	-(r)are-	kak-are-	tabe-rare-
	丁寧	-(i)mas-	kak-imas-	tabe-mas-
	可能	-e-	kak-e-	-
	否定	-(a)na-	kak-ana-	tabe-na-
	希望	-(i)ta-	kak-ita-	tabe-ta-
統語接尾辞	非完了終止	-(r)u	kak-u	tabe-ru
	完了終止	-(i)ta	ka-ita	tabe-ta
	前望肯定	-(y)ou	kak-ou	tabe-you
	完了	-(i)te	ka-ite	tabe-te
	否定	-(a)zu	kak-azu	tabe-zu
	仮定条件	-(r)eba	kak-eba	tabe-reba
	命令	-e/-ro	kak-e	tabe-ro

派生接尾辞・統語接尾辞

	役割	接尾辞	用例	
			子音幹	母音幹
派生接尾辞	使役	-(s)ase-	kak-ase-	tabe-sase-
	受身・尊敬・可能	-(r)are-	kak-are-	tabe-rare-
	丁寧	-(i)mas-	kak-imas-	tabe-mas-
	可能	-e-	kak-e-	-
	否定	-(a)na-	kak-ana-	tabe-na-
	希望	-(i)ta-	kak-ita-	tabe-ta-
統語接尾辞	非完了終止	-(r)u	kak-u	tabe-ru
	完了終止	-(i)ta	ka-ita	tabe-ta
	前望肯定	-(y)ou	kak-ou	tabe-you
	完了	-(i)te	ka-ite	tabe-te
	否定	-(a)zu	kak-azu	tabe-zu
	仮定条件	-(r)eba	kak-eba	tabe-reba
	命令	-e/-ro	kak-e	tabe-ro

不規則変化

不規則変化　音便形

完了 -(i)ta, -(i)te, など

語幹末子音		語幹末子音		語例	備考
-k	-(i)ta	-ϕ	-ita	書いた	イ音便
-g	-(i)ta	-ϕ	-ida	泳いだ	イ音便+連濁
-r	-(i)ta	-t	-ϕta	切った	促音便
-t	-(i)ta			立った	
-w	-(i)ta			買った	
-b	-(i)ta	-n’	-ϕda	飛んだ	撥音便+連濁
-n	-(i)ta			死んだ	
-m	-(i)ta			読んだ	
-s	-(i)ta	-s	-ita	貸した	音便変化なし

不規則変化　音便形

完了 -(i)ta, -(i)te, など

語幹末子音		語幹末子音		語例	備考
-k	-(i)ta	-φ	-ita	書いた	イ音便
-g	-(i)ta	-φ	-ida	泳いだ	イ音便+連濁
-r	-(i)ta	-t	-φta	切った	促音便
-t	-(i)ta			立った	
-w	-(i)ta			買った	
-b	-(i)ta	-n’	-φda	飛んだ	撥音便+連濁
-n	-(i)ta			死んだ	
-m	-(i)ta			読んだ	
-s	-(i)ta	-s	-ita	貸した	音便変化なし
ik	-(i)ta	it	-φta	行った	例外
tow	-(i)ta	toφ	-uta	問うた	例外

形態素解析システム MAJO

Morphological Analyzer of Japanese based On Derivational Grammar

- 派生文法に基づく形態素解析

形態素文法への変更

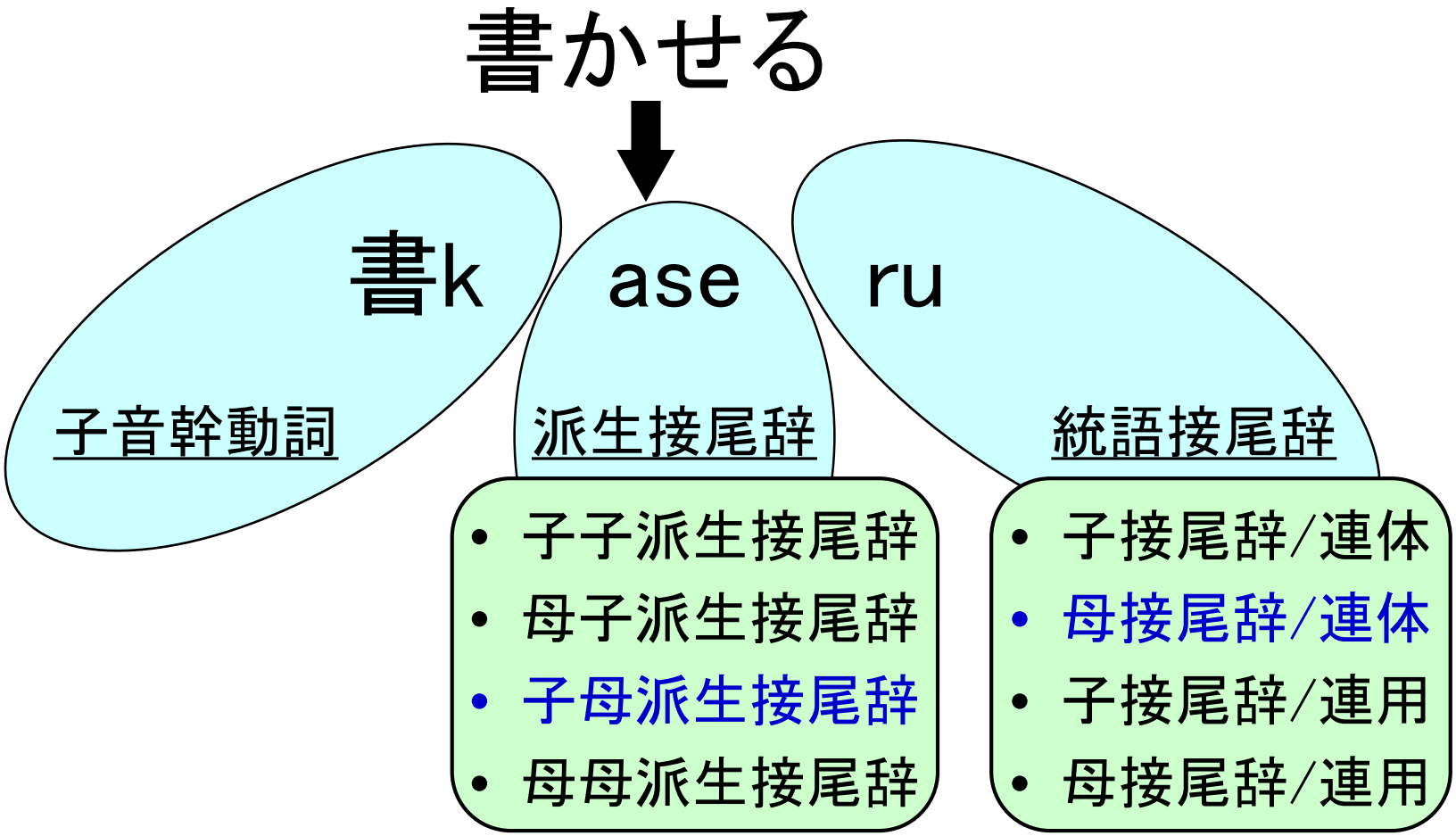
書かせる

⬇

書kaseru

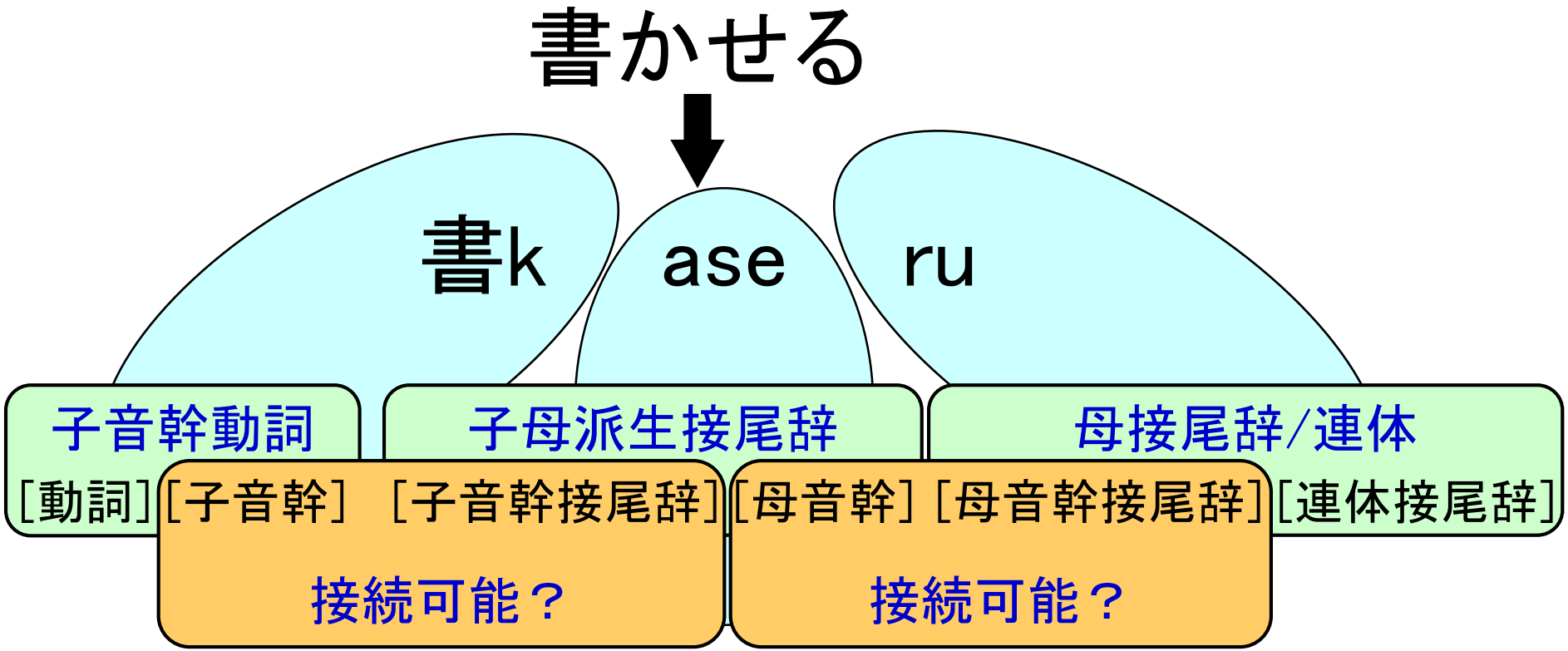
入力文を漢字ローマ字混じり文に変換

形態素文法への変更



品詞(接尾辞)を接続規則によって細かく分類

形態素文法への変更



品詞に左連接属性と右連接属性を付与

形態素文法を連接属性間の接続可能性で与える

形態素文法への変更

右連接属性＼左連接属性	名詞	動詞	格接尾辞	名詞接尾辞	母音幹接尾辞	子音幹接尾辞	接頭辞	句読点
子音幹	－	－	－	－	－	o	－	－
母音幹	o	o	o	－	o	－	o	o

子音幹動 … 辞/連体

[動詞] [子音幹] [子音幹接尾辞] [母音幹] [母音幹接尾辞] [連体接尾辞]

接続可能○

接続可能○

品詞に左連接属性と右連接属性を付与

形態素文法を連接属性間の接続可能性で与える

接続行列

左連接属性 / 右連接属性	名詞	動詞	格接尾辞	名詞接尾辞	母音幹接尾辞	子音幹接尾辞	接頭辞	句読点
子音幹	−	−	−	−	−	o	−	−
母音幹	o	o	o	−	o	−	o	o
連体接尾辞	o	o	o	−	−	−	o	o
連用接尾辞	o	o	−	−	−	−	o	o
名詞幹	o	o	o	o	−	−	o	o
格接尾辞	o	o	o	−	−	−	o	o

接続行列　　接続コストの付加

左連接属性 / 右連接属性	名詞	動詞	格接尾辞	名詞接尾辞	母音幹接尾辞	子音幹接尾辞	接頭辞	句読点
子音幹	–	–	–	–	–	5	–	–
母音幹	20	20	20	–	5	–	20	20
連体接尾辞	10	20	25	–	–	–	20	20
連用接尾辞	20	20	–	–	–	–	20	20
名詞幹	15	20	5	5	–	–	20	20
格接尾辞	10	10	30	–	–	–	20	20

接続コストの和が最小となる組合せを解とする

不規則変化への対応

異形態(音便変化形)の登録

語幹末尾	異形態	備考
-k	書i -ta	イ音便
-g	泳i -da	イ音便+連濁
-r	切t -ta	促音便
-t	立t -ta	
-w	買t -ta	
-b	飛n' -da	撥音便+連濁
-n	死n' -da	
-m	読n' -da	
-s	貸s -ita	音便変化なし
ik	行t -ta	例外
tow	問u -ta	例外

形態素解析システムMAJO

長所

① 動詞の語形変化に対する処理が単純

② 品詞の種類が少なく、文法が単純

③ 口語表現への対処が容易

長所①　単純な処理

学校文法による解析

書か	力行五段活用動詞未然形
せ	助動詞未然形
られ	助動詞連用形
ます	助動詞終止形

- 動詞の細かな分類が必要
- 活用形のチェックが必要

MAJOによる解析

書k-	子音幹動詞
-ase-	子母派生接尾辞
-rare-	母母派生接尾辞
-mas-	母子派生接尾辞
-u	子統語接尾辞

- 動詞の分類は2種類だけ
- 活用形のチェックは不要

長所② 単純な文法

左連接属性 / 右連接属性	名詞	動詞	格接尾辞	名詞接尾辞	母音幹接尾辞	子音幹接尾辞	接頭辞	句読点
子音幹	–	–	–	–	–	5	–	–
母音幹	20	20	20	–	5	–	20	20
連体接尾辞	10	20	25	–	–	–	20	20
連用接尾辞	20	20	–	–	–	–	20	20
名詞幹	15	20	5	5	–	–	20	20
格接尾辞	10	10	30	–	–	–	20	20

システム	接続行列
MAJO	24 x 26
JUMAN	195 x 165

MAJOの形態素文法は単純

長所③　　口語表現への対応

- ら抜き言葉

食be-rare-ru	通常の表現
食be-re-ru	慣用的表現

可能の派生接尾辞 -re- の追加

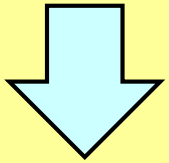
- 使役「さす」

書k-ase-ru	通常の表現
書k-as-u	慣用的表現

使役の派生接尾辞 -(s)as- の追加

短所　　解析対象文字列の増加

書かせられます　　7文字



書kaseraremasu　　13文字

- 辞書引き回数の増加
- 解候補の組合せの増加

評価実験

- 実験対象
  - EDR日本語コーパス中の1,000文
  - 出力は一つだけ

実験結果

システム	コーパス	文数	形態素数	誤り数	エラー率
MAJO	EDR	1,000	25,012	469	1.88%
渕ら[’96]	EDR	10,000	207,574	2,040	0.98%
丸山ら[’92]	新聞記事	1,016	29,024	687	2.36%

単純な文法で十分な精度

日本語ーウイグル語機械翻訳

日本・ウイグル・ウズベキスタン

日本・ウイグル・ウズベキスタン

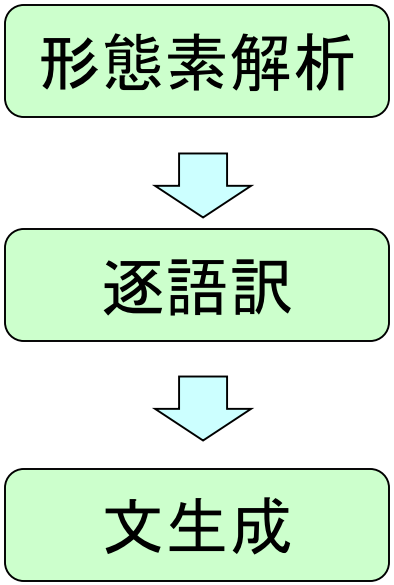
日本語－ウイグル語翻訳

日本語とウイグル語

- 共に膠着語
- 構文的類似性が高い
- 形態論的類似性も高い　[小川ら’00]

形態素解析後、逐語訳による翻訳

日本語－ウイグル語機械翻訳



肉をたくさん食べた。

肉　を　たくさん　食べ　た　。

gosh　–ni　jiq　yé　-dim　.

goshni jiq yédim.

日本語－ウイグル語機械翻訳

共通点
- 格助詞
- 主語の省略
- 複合名詞

相違点
- 人称接尾辞
- 動詞の活用

肉をたくさん食べた。

肉　を　たくさん　食べ　た　。

↓　↓　↓　↓　↓　↓

gosh　–ni　jiq　yé　-dim　.

goshni jiq yédim.

派生文法の導入

- ウイグル語は活用しない
- 派生文法によれば、日本語も活用しない

派生文法を用いた動詞句翻訳

書kareta

書k -(r)are -(i)ta

yaz -(i)l -di

yazildi

作rareta

作r -(r)are -(i)ta

yasa -(i)l -di

yasaldi

1. 逐語訳による動詞句翻訳

派生文法を用いた動詞句翻訳

書kareta

書k -(r)are -(i)ta

yaz -(i)l -di

yazildi

作rareta

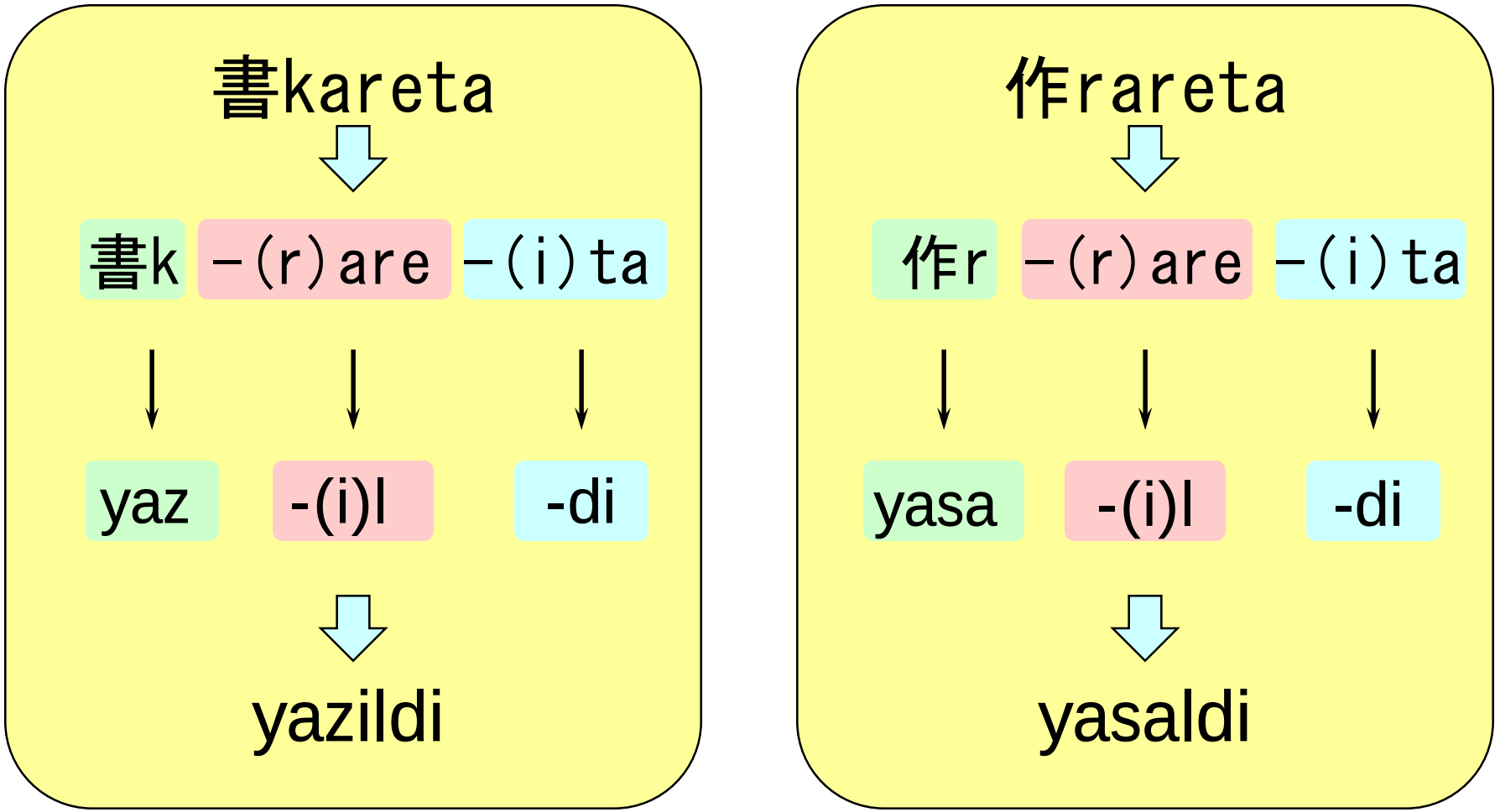
作r -(r)are -(i)ta

yasa -(i)l -di

yasaldi

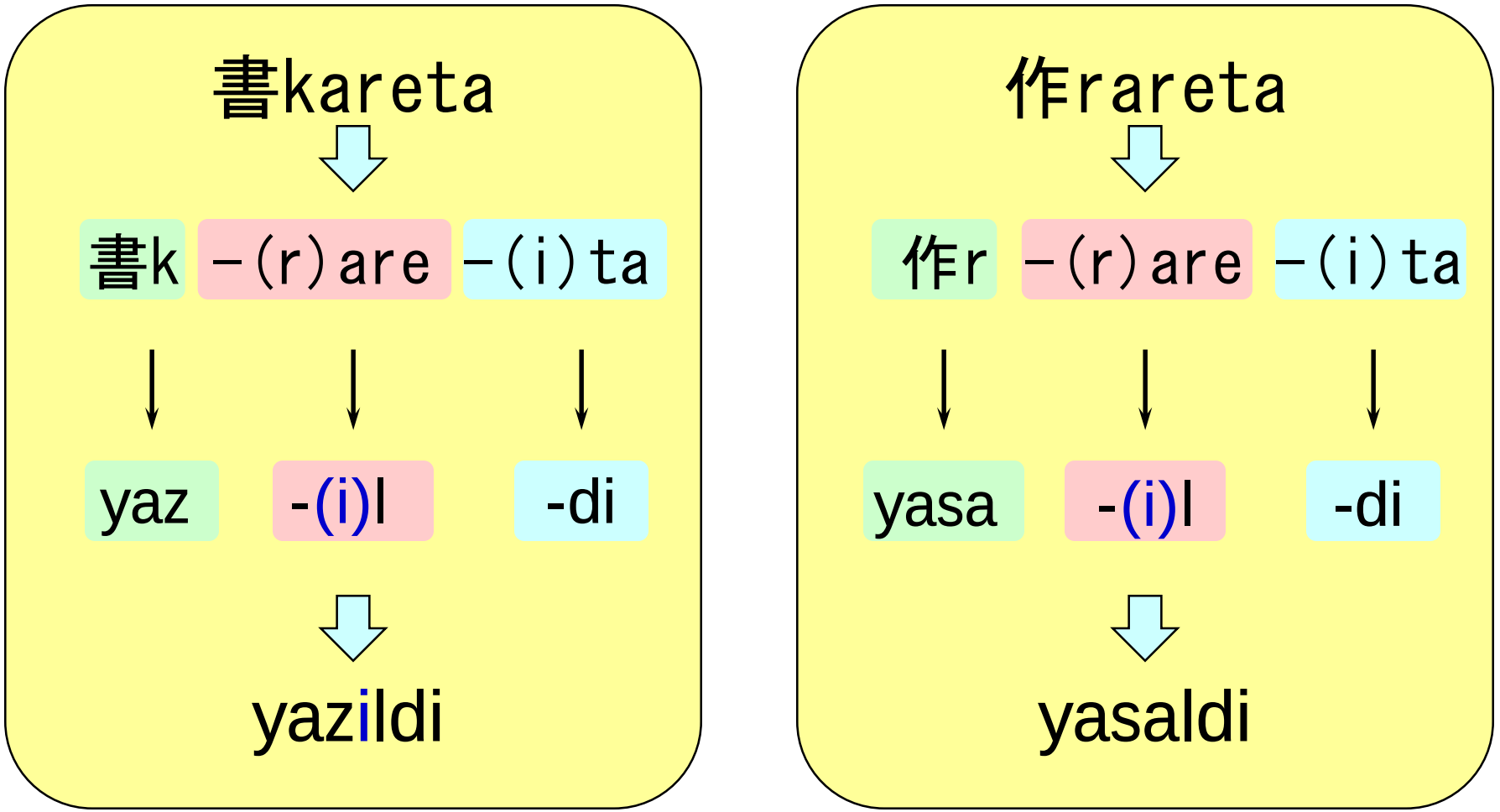
2. 子音幹と母音幹の区別

派生文法を用いた動詞句翻訳



3. 派生接尾辞・統語接尾辞の存在

派生文法を用いた動詞句翻訳



4. 連結母音・連結子音の欠落規則

問題点

単純な逐語訳では不自然な翻訳となる例

① 終止形と連体形

② 派生語幹の不一致

③ サ変動詞

問題点

単純な逐語訳では不自然な翻訳となる例

① 終止形と連体形

② 派生語幹の不一致

③ サ変動詞

問題点①　終止形と連体形

終止形

彼	が	書	いた	。
↓	↓	↓	↓	↓
U	-	yaz-	-di	.

U yazdi.

連体形

彼	が	書	いた	本
↓	↓	↓	↓	↓
U	-	yaz-	-ghan	kitap

U yazghan kitap

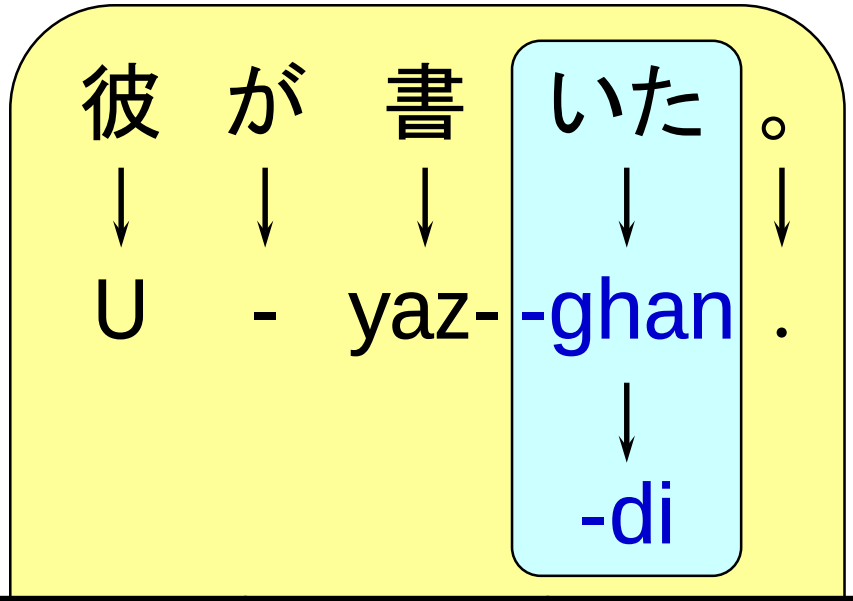
ウイグル語では終止形と連体形で接尾辞が異なる

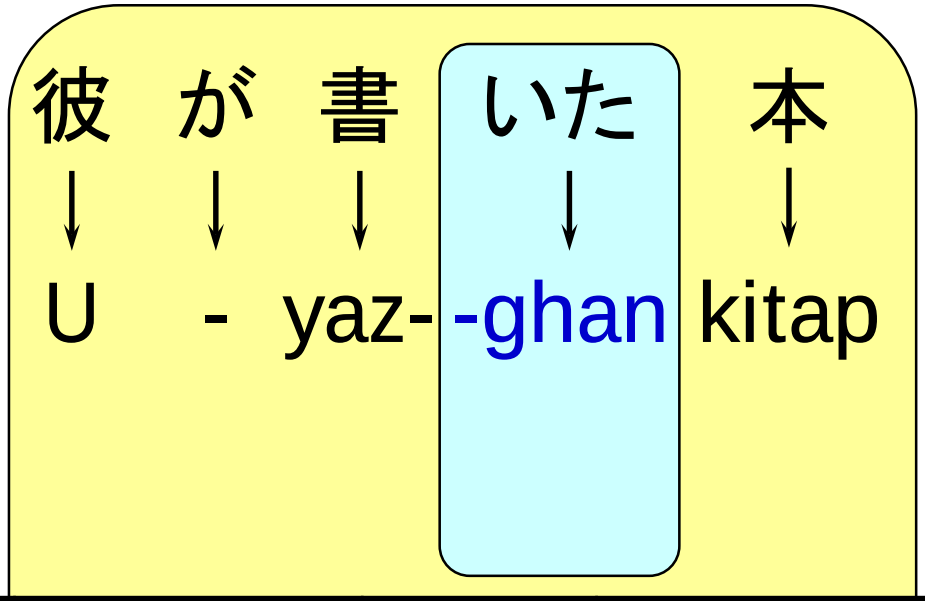
訳語置換表

前後の単語が条件を満たす → 訳語を置換

日本語	基本訳語	前接 ウイグル語	後接 ウイグル語	新訳語	新品詞
−(i)ta	-ghan	*	文末	-di	終止接尾辞
		*	句読点	-di	終止接尾辞
		*	終助辞	-di	終止接尾辞
−katta	-ken	-ma	*	-ghan	連体接尾辞
登録	tizimlash	*	kil	tizimla	サ変動詞
si, su, se	, kil	サ変動詞	*	-	動詞

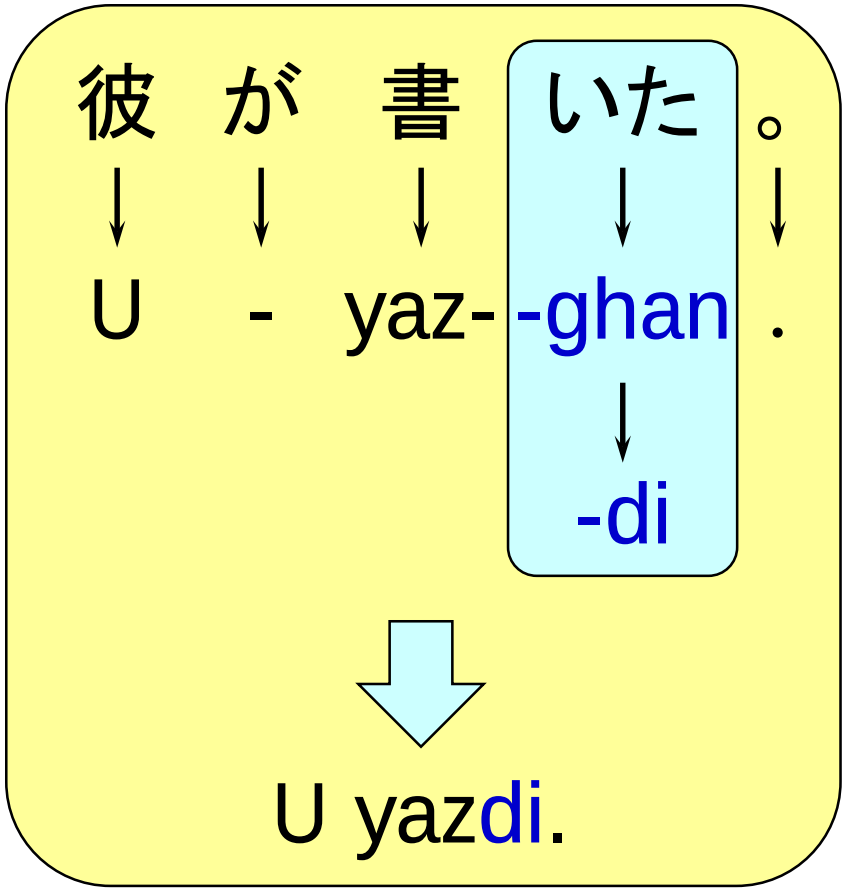
翻訳例

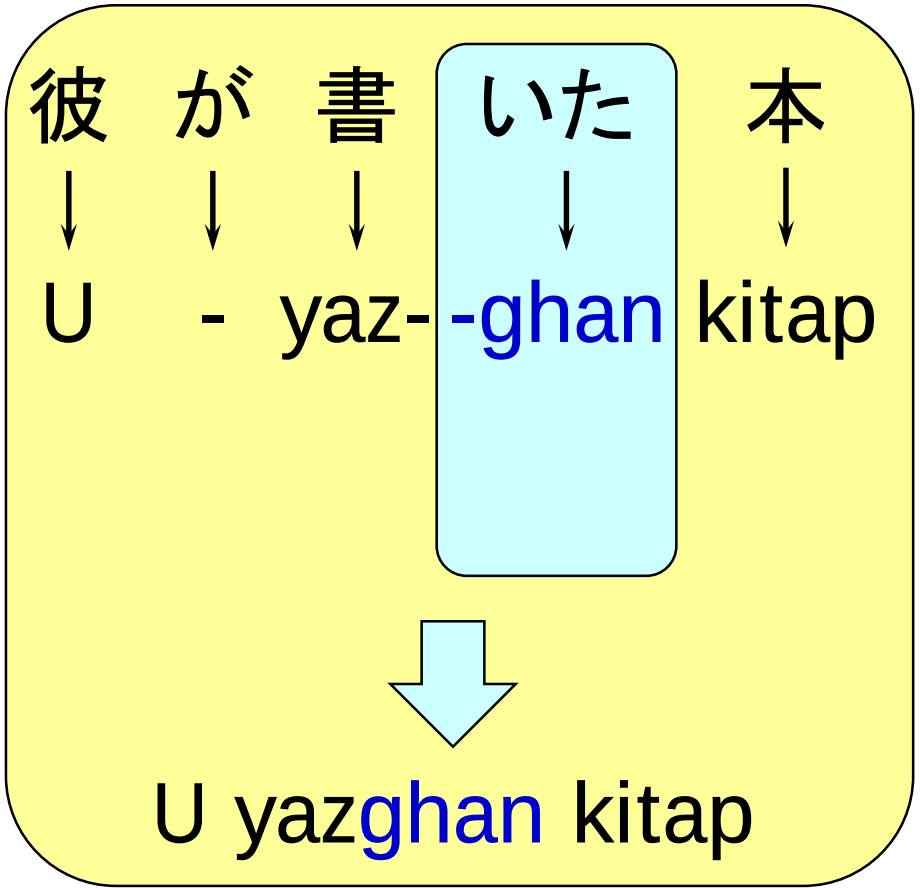




日本語	基本訳語	前接ウイグル語	後接ウイグル語	新訳語	新品詞
−(i)ta	-ghan	*	文末	-di	終止接尾辞
		*	句読点	-di	終止接尾辞
		*	終助辞	-di	終止接尾辞

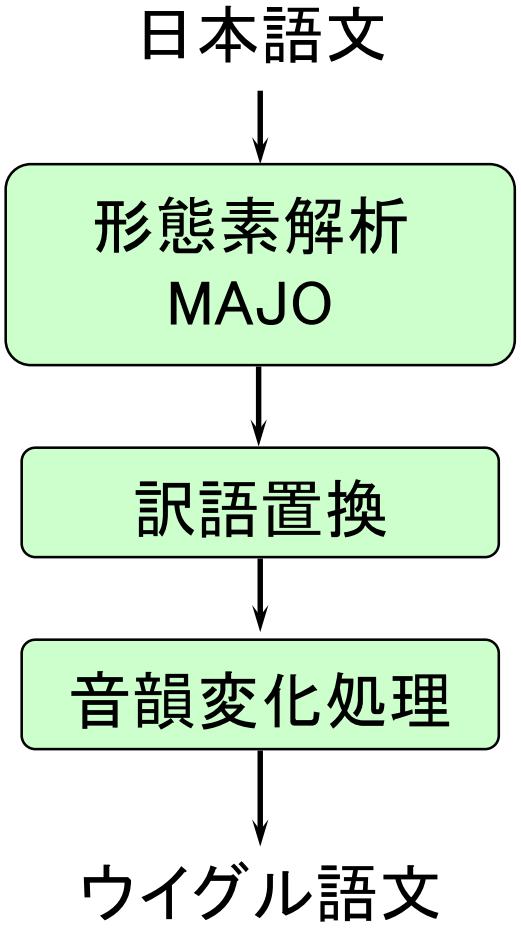
翻訳例


彼 が 書 いた 。
U - yaz- -ghan .
-di
U yazdi.


彼 が 書 いた 本
U - yaz- -ghan kitap
U yazghan kitap

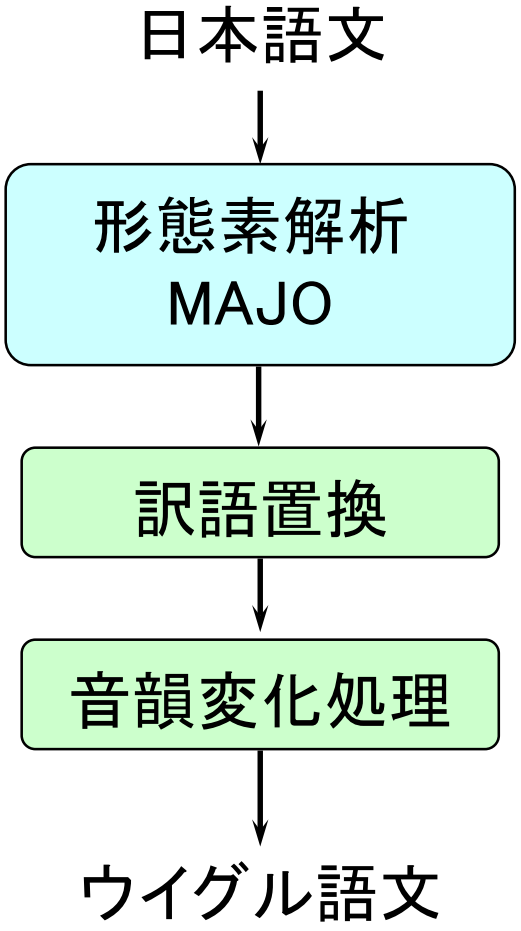
より自然な翻訳

翻訳システムの構成



書かせられました。

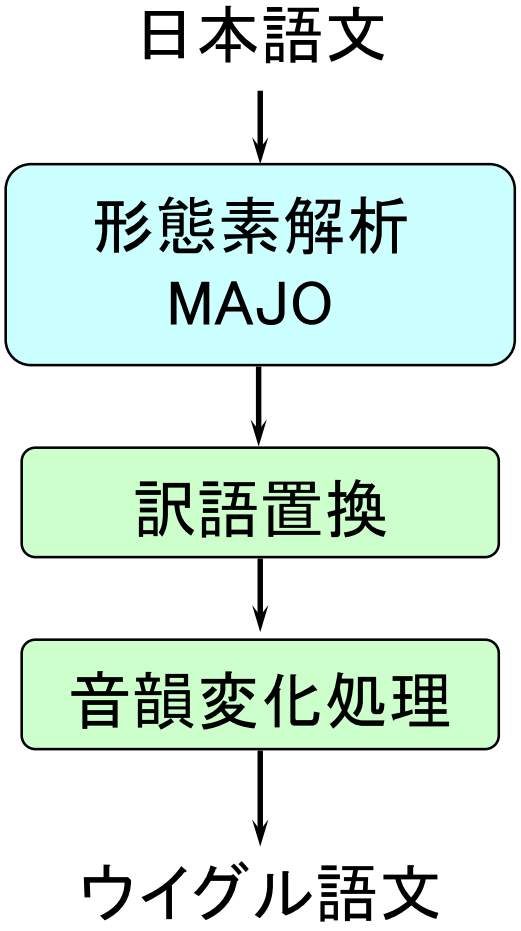
翻訳システムの構成



書かせられました。

日本語	品詞	ウイグル語
書k	子音動詞	yaz
ase	派生接尾辞	-ghuz
rare	派生接尾辞	-(i)l
mas	派生接尾辞	-
ita	統語接尾辞	-ghan
。	句読点	.

翻訳システムの構成

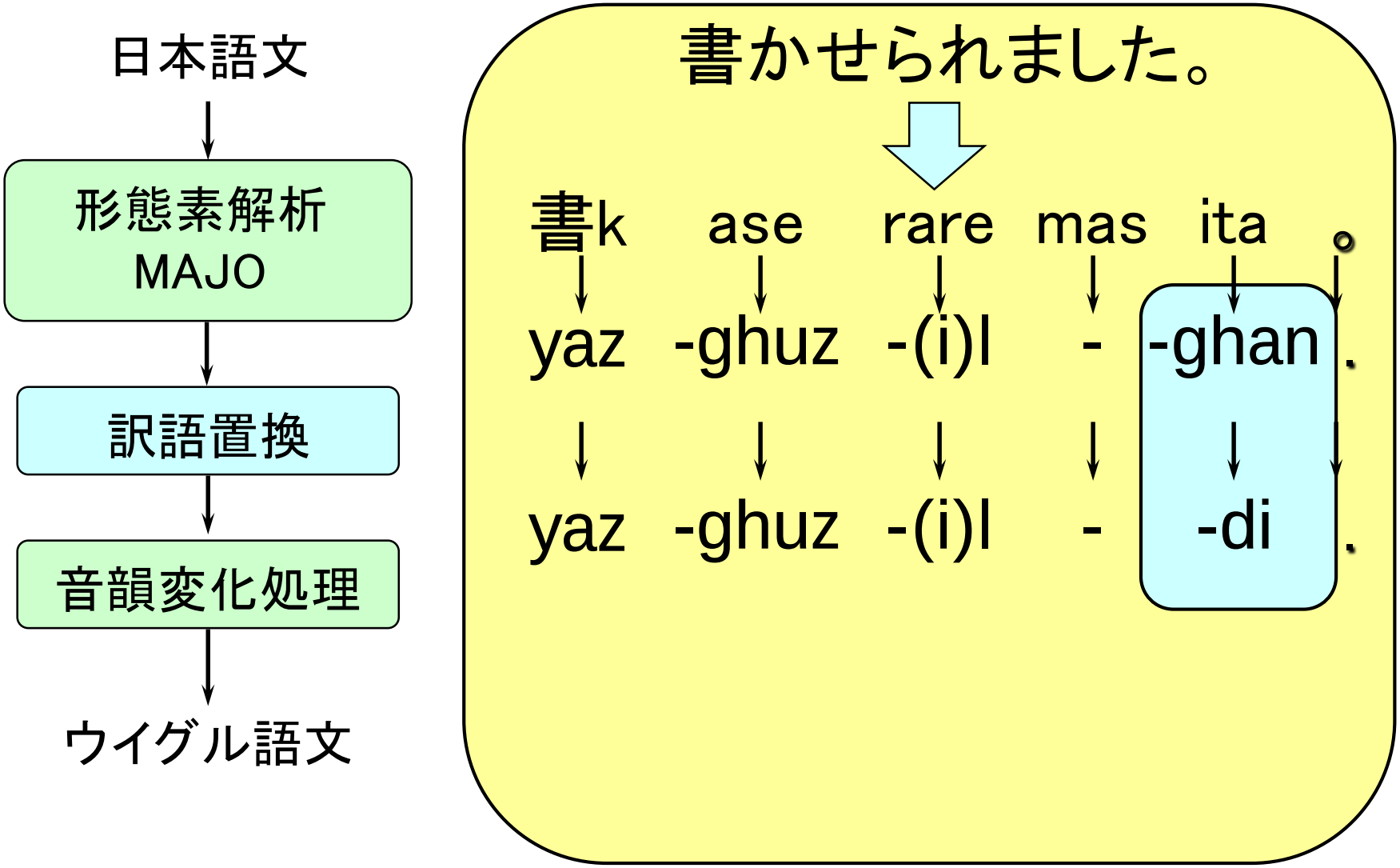


書かせられました。

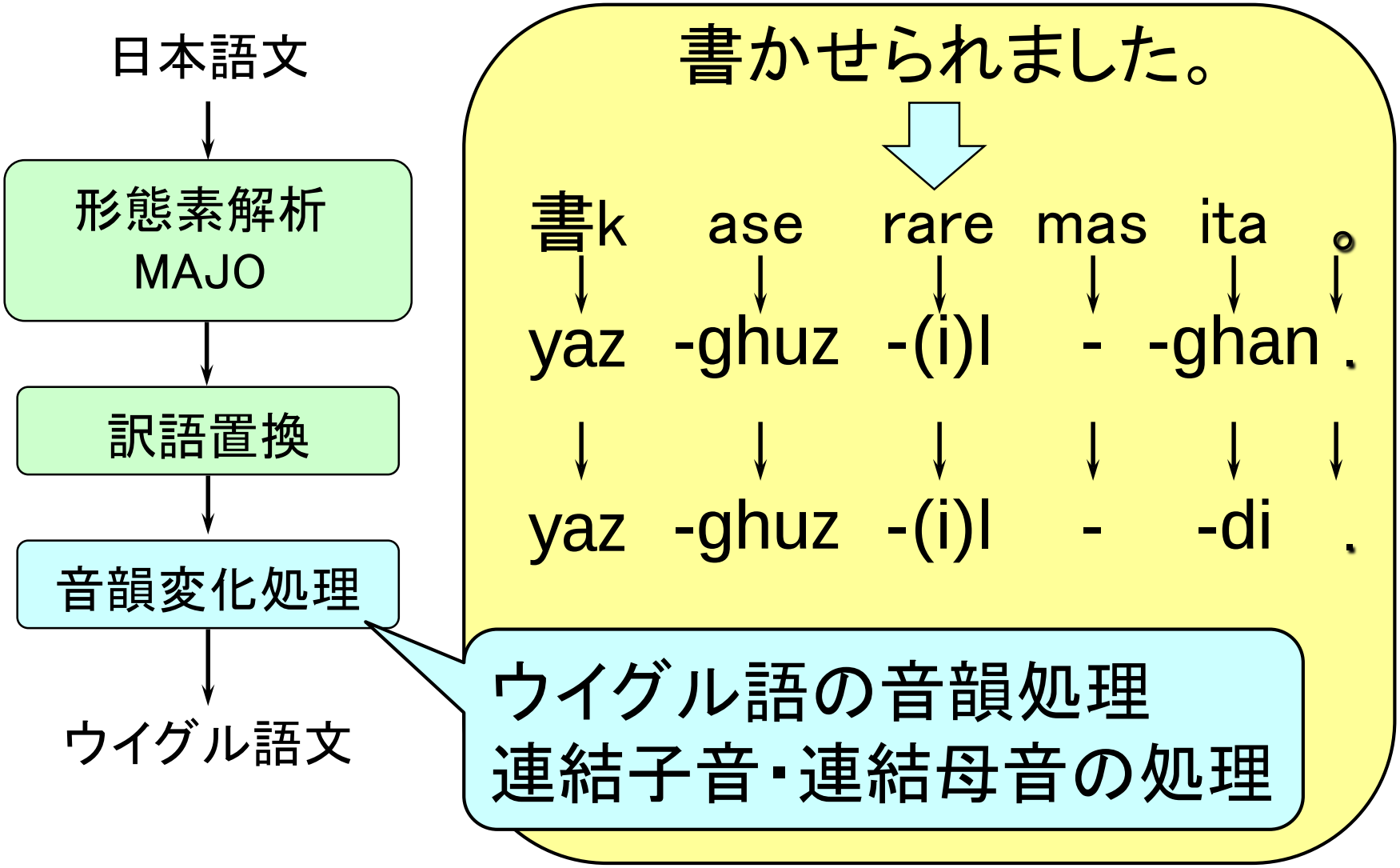
書k　ase　rare　mas　ita　。

yaz　-ghuz　-(i)l　-　-ghan　.

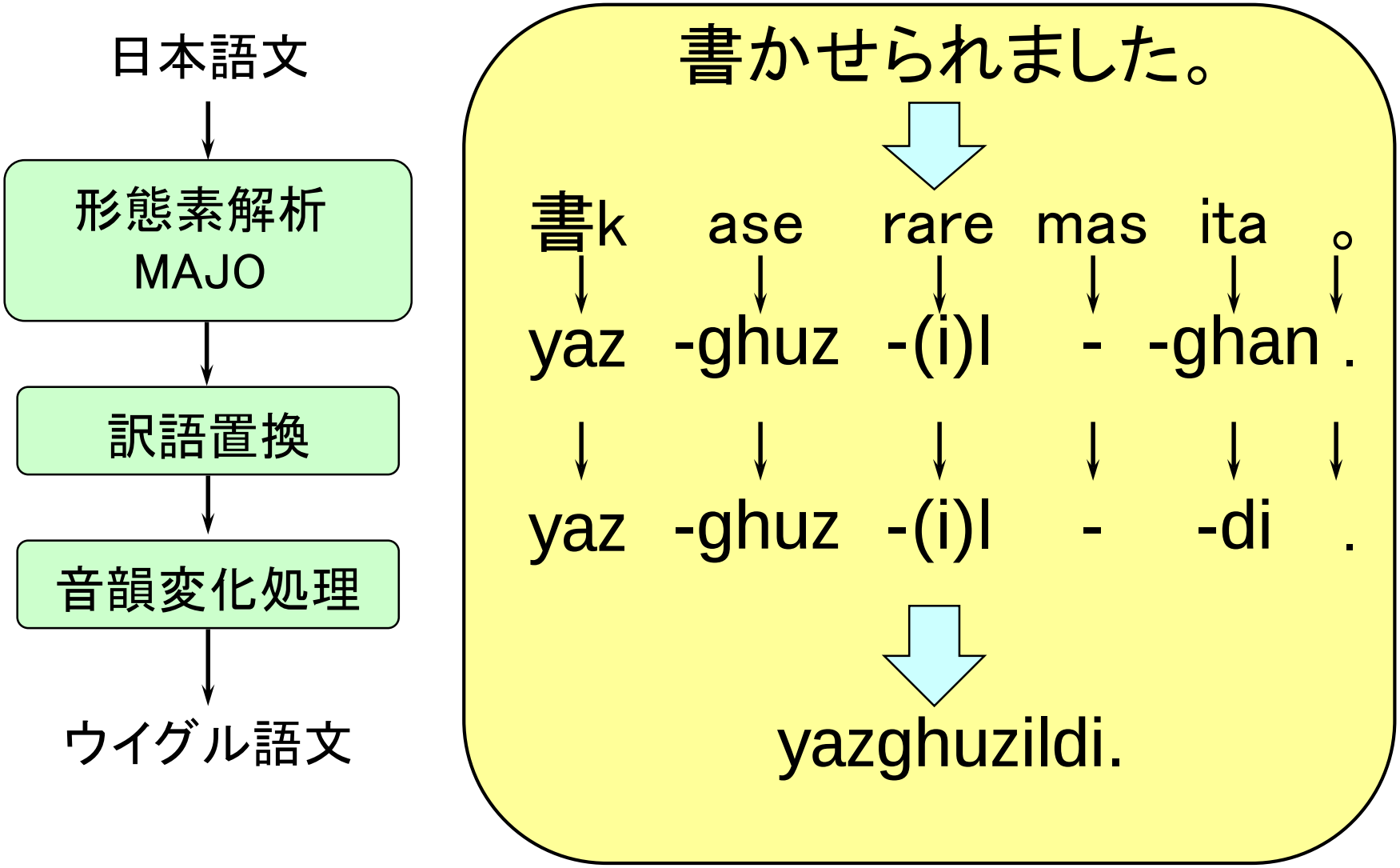
翻訳システムの構成


日本語文
形態素解析
MAJO
訳語置換
音韻変化処理
ウイグル語文
書かせられました。
書k ase rare mas ita 。
yaz -ghuz -(i)l - -ghan .
yaz -ghuz -(i)l - -di .

翻訳システムの構成


日本語文
形態素解析
MAJO
訳語置換
音韻変化処理
ウイグル語文
書かせられました。
書k ase rare mas ita 。
yaz -ghuz -(i)l - -ghan .
yaz -ghuz -(i)l - -di .
ウイグル語の音韻処理
連結子音・連結母音の処理

翻訳システムの構成


日本語文
形態素解析
MAJO
訳語置換
音韻変化処理
ウイグル語文
書かせられました。
書k ase rare mas ita 。
yaz -ghuz -(i)l - -ghan .
yaz -ghuz -(i)l - -di .
yazghuzildi.

ウイグル語ー日本語電子化辞書

飯沼英三:
「ウイグル語辞典」

- 収録語数:約16,000語
- 修正4年がかり

品詞	見出し数	日本語への平均対訳数
名詞	7,259	1.86
動詞	3,682	2.36
形容詞	2,868	2.12
動作名詞	1,046	2.12
副詞	691	2.09
助数詞	142	1.61
感嘆詞	48	2.14
代名詞	19	1.79
その他	33	2.59
合計	15,788	2.05

日本語－ウイグル語電子辞書

逆向きに変換

- 収録語数：約20,000語
- 修正に2年
- 機械翻訳システムの辞書として利用

品詞	見出し数	ウイグル語への平均対訳数
名詞	8,457	1.51
動詞	5,411	1.60
形容詞	3,480	1.68
サ変名詞	1,785	1.31
副詞	784	1.76
助数詞	156	1.47
感嘆詞	73	1.71
接続詞	20	1.55
合計	20,166	1.56

動詞句翻訳実験

システム
- MAJO　辞書: 20,000語中
  動詞:　5,400語
- 訳語置換
- ウイグル語整形

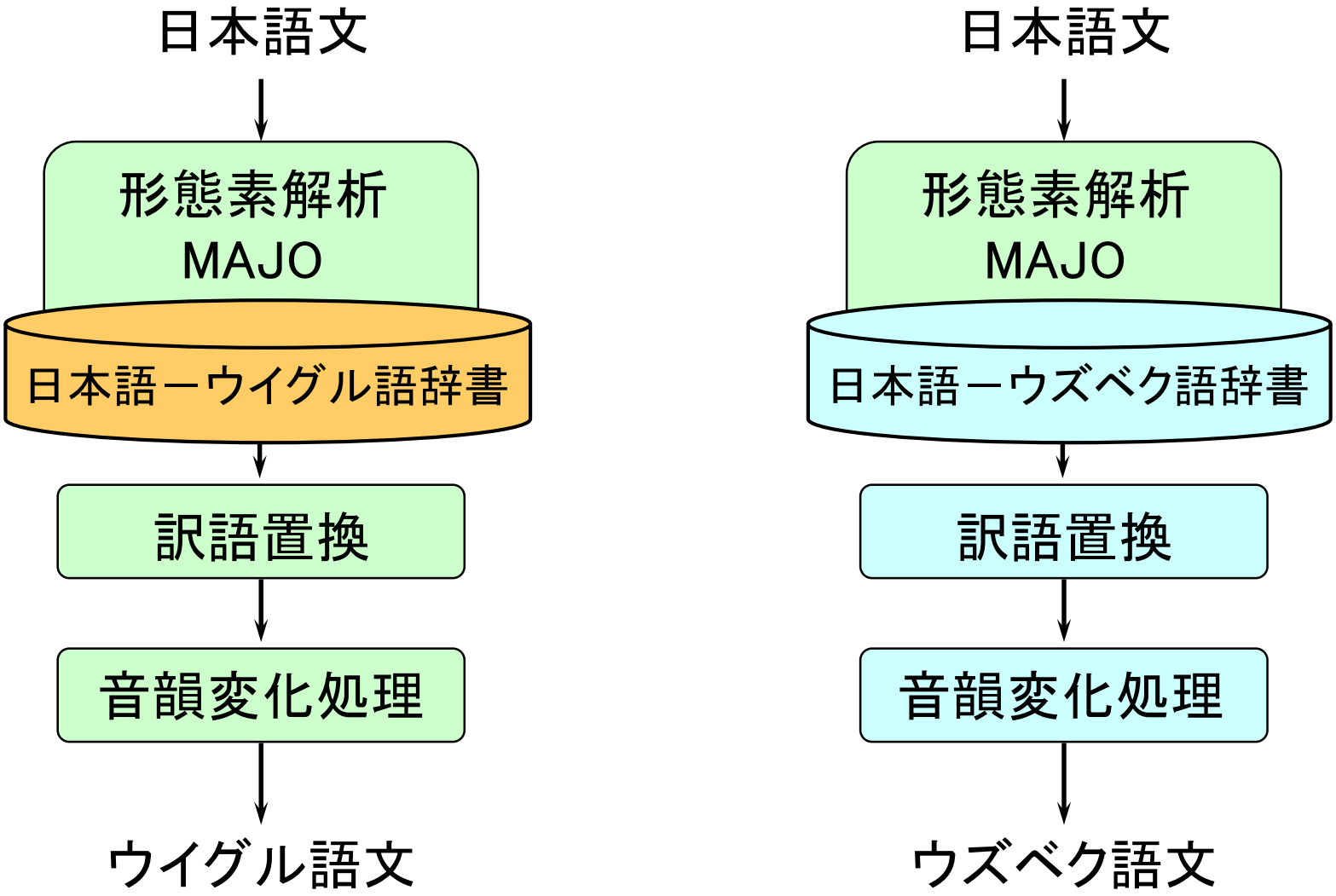
データ
- 新聞の社説3編（環境破壊）
  - 文数:　136
  - 動詞句数:306

実験結果

	誤り原因	個数	割合
自然な翻訳		212	69.3%
不自然だが意味は分かる	終止形と連体形の区別	3	23.2%
	音韻規則の適用	8	
	訳語の多様性	55	
	認証接尾辞の選択	5	
意味が理解できない	直訳不能	21	7.5%
	形態素解析	3	

日本語ーウズベク語機械翻訳

日本語－ウズベク語機械翻訳



ウイグル語とウズベク語の語彙

日本語	ウイグル語	ウズベク語
雲	bulut	bulut
噴水	fontan	fontan
皇帝	impérator	imperator
恐怖	dehshet	dahshat
頭	bash	bosh
ある	bar	bor
育てる	östürmek	ochirmoq
雇う	yallimaq	yollamoq

ウイグル語とウズベク語の相違

母音調和

前母音	e ü ö
後母音	a u o
中立	i é

östür

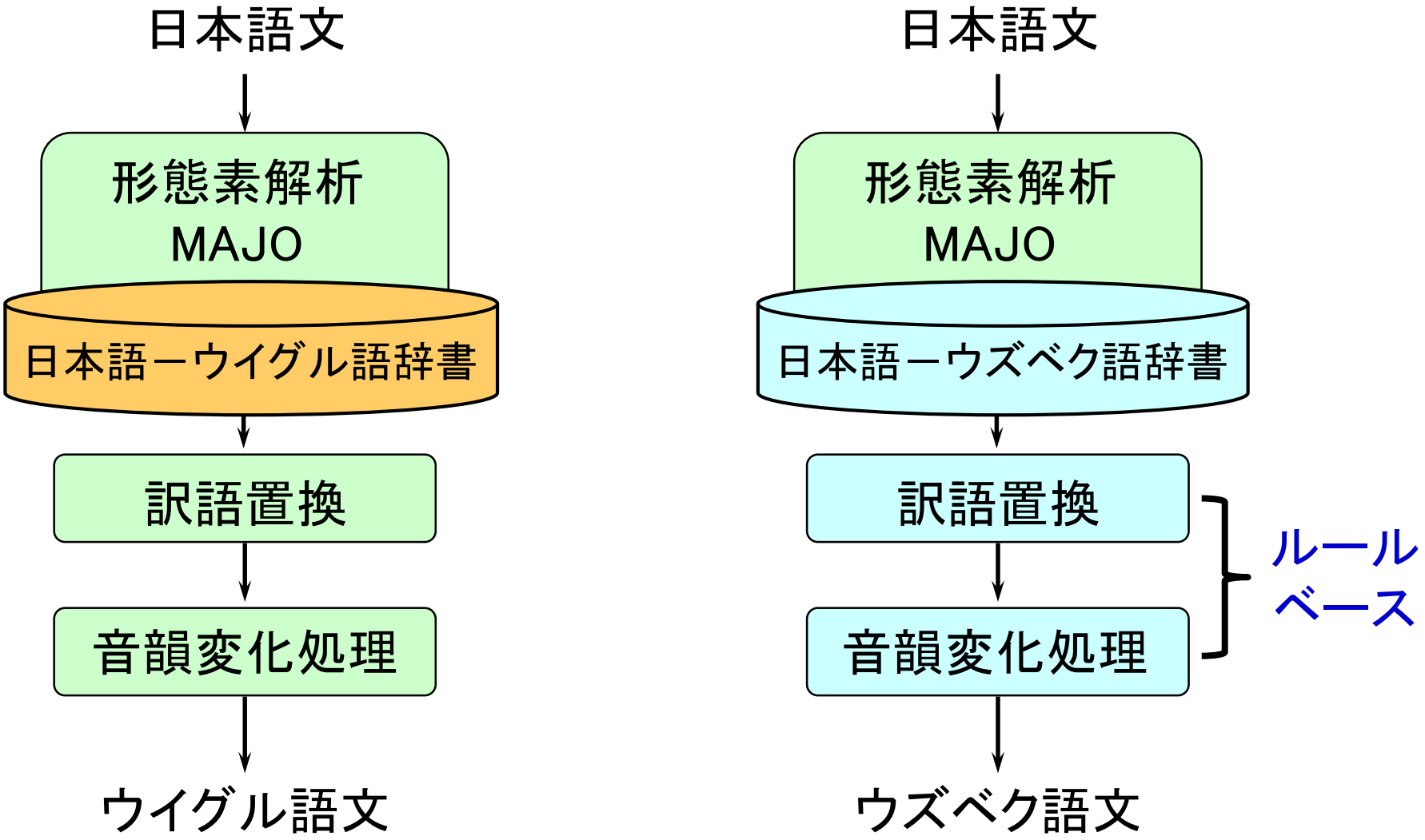
yallimaq

mek

その他の音韻規則

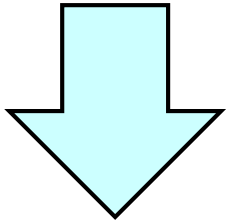
育てる	östürmek	ochirmoq
雇う	yallimaq	yollamoq

日本語－ウズベク語機械翻訳



統計的手法の導入

- ルールベースとの融合
  - 訳語選択
  - 音韻変化処理
  - ウズベク語辞書との相互変換



ウイグル語（対訳）コーパスの構築

まとめ

- 日本語形態素解析
  - 派生文法
  - 派生文法に基づく形態素解析システム
- 日本語－ウイグル語機械翻訳
  - 派生文法に基づく翻訳
  - 実験
- 日本語－ウズベク語機械翻訳

おまけ

文法的に正しい「ら抜き言葉」

- 学校文法における「ら抜き言葉」
  - 文法違反
  - 学校文法では扱えない

- 派生文法における扱い
  - 可能の派生接尾辞を再考してみる

学校文法の助動詞活用表

種類	受身・尊敬・自発・可能		使役		
語	れる	られる	せる	させる	しめる
未然形	れ	られ	せ	させ	しめ
連用形	れ	られ	せ	させ	しめ
終止形	れる	られる	せる	させる	しめる
連体形	れる	られる	せる	させる	しめる
仮定形	れれ	られれ	せれ	させれ	しめれ
命令形	れよ れろ	られよ られろ	せよ せろ	させよ させろ	しめよ しめろ
接続	五段・サ変の未然形	その他の未然形	五段・サ変の未然形	その他の未然形	用言の未然形

「食べれる」は認められない

派生文法の接尾辞

	役割	接尾辞	用例	
			子音幹	母音幹
派生接尾辞	使役	-(s)ase-	kak-ase-	tabe-sase-
	受身・尊敬・可能	-(r)are-	kak-are-	tabe-rare-
	丁寧	-(i)mas-	kak-imas-	tabe-mas-
	可能	-e-	kak-e-	-
	否定	-(a)na-	kak-ana-	tabe-na-
	希望	-(i)ta-	kak-ita-	tabe-ta-
統語接尾辞	非完了終止	-(r)u	kak-u	tabe-ru
	完了終止	-(i)ta	ka-ita	tabe-ta
	前望肯定	-(y)ou	kak-ou	tabe-you
	完了	-(i)te	ka-ite	tabe-te
	否定	-(a)zu	kak-azu	tabe-zu
	仮定条件	-(r)eba	kak-eba	tabe-reba
	命令	-e/-ro	kak-e	tabe-ro

派生文法の接尾辞

	役割	接尾辞	用例	
			子音幹	母音幹
派生接尾辞	使役	-(s)ase-	kak-ase-	tabe-sase-
	受身・尊敬・可能	-(r)are-	kak-are-	tabe-rare-
	丁寧	-(i)mas-	kak-imas-	tabe-mas-
	可能	-(r)e-	kak-e-	tabe-re
	否定	-(a)na-	kak-ana-	tabe-na-
	希望	-(i)ta-	kak-ita-	tabe-ta-
統語接尾辞	非完了終止	-(r)u	kak-u	tabe-ru
	完了終止	-(i)ta	ka-ita	tabe-ta
	前望肯定	-(y)ou	kak-ou	tabe-you
	完了	-(i)te	ka-ite	tabe-te
	否定	-(a)zu	kak-azu	tabe-zu
	仮定条件	-(r)eba	kak-eba	tabe-reba
	命令	-e/-ro	kak-e	tabe-ro

可能の派生接尾辞 -(r)e-

書k-e-ru

食be-re-ru　　ら抜き言葉

接尾辞の例外でなくなる

- 文法が単純になる
- 実際に多くの人が使用している

「ら抜き言葉」は文法的に正しい

学校文法における可能

- 「れる」「られる」で表現
  - 「食べられる」
  - 「書かれる」

実際には言わない

- 可能動詞
  - 「書ける」

なぜ可能だけ特別なのか？

「ら抜き言葉」を扱えないのは学校文法の欠点

真面目な話

- 「ら抜き言葉」は方言の一種
  - 「ら抜き言葉」が文法違反なら関西弁も違反
- 言語は文法が単純になる方向へ変化
  - 例外が減っていく