Apertium based Automatic Interlinear Gloss Generator

Alex Shi
wshi2@swarthmore.edu

01. Introduction

Interlinear gloss provides a syntactic and semantic description of text passages. The purpose of this project is to create a script that will automatically gloss text according to the Lipzig rules using an Apertium morphological analyser and bilingual translator.

02. Usage

To use the autoglosser, enter the following:

```
echo "your sentence here" | python3 autoglosser.py
--igtdef <igt file> --inputlang <ISO code> --glosslang
<ISO code> --biltrans <bilingual transducer directory>
```

The igt file should have 5 columns, which are, from left to right: lemma, part of speech, morphological tags, prefixes, and suffixes. The first three columns describe the context, and the last two columns describe how it should be glossed.

For example:

lemma	POS	tags	pref	suff
-	n	pl	-	-PL
sheep	n	pl	-	[PL]
-	-	perf	-	.PFV
-	-	p3.sg	3SG-	-

The first line says that when we have a plural noun, it should be glossed with a “-PL” suffix. The second line is more specific, and says that if we have “sheep<n><pl>”, then it should be glossed with a “[PL]” instead. The autoglosser should prefer using the more specific glossing rules if they exist.
The third line says that perfective words should be tagged with “.PFV”, the standard interlinear gloss abbreviation.
Finally, the fourth line says that words tagged with both third person and singular should get a “3SG” prefix.

03. Evaluation

I have created a (very limited) Kyrgyz igt file to test the autoglosser.
Here are its results compared to the gloss made by Professor Washington:

```
Азамат менен Айгүл бакчада.
JW: Azamat[NOM] and Aygül[NOM] garden-LOC[COP.NPST.3]
AG: Azamat.M[NOM] and Aygül.F[NOM] garden-LOC [be.NPST.3]

Аба ырайы бүгүн абдан жакшы, жылуу.
JW: weather[NOM] today very good, warm[COP.NPST.3]
AG: weather[NOM] today very good [be.NPST.3] warm [be.NPST.3]

Бирок кечээ өтө суук эле!
JW: but yesterday very cold cop-PST.DEF[3]
AG: but yesterday very cold be[3SG]-PST.DEF

Ушул себептен алар эшикте ойной алган жок.
JW: this reason-ABL they outside-LOC play-VADV can-PST-NEG[3]
AG: this reason-ABL prpers[3PL.NOM] door-LOC play-VADV can[3SG]-NEG-PST.DEF

Азамат менен Айгүл ойногонду жакшы көрөт,
JW: Azamat[NOM] and Aygül[NOM] play-VN-ACC like-NPST-3
AG: azamat.M[NOM] and aygül.F[NOM] play-VN-ACC like-3SG-NPST

алар дайыма чоң үйдүн алдындагы бакчада чогуу ойношот.
JW: they always big house-GEN front-POSS.3-LOC.ATTR garden-LOC together play-PL-NPST-3
AG: prpers[3PL.NOM] always big home-GEN front-POSS.3-LOC.ATR garden-LOC together play-3PL-NPST
```

The autoglosser is able to produce gloss similar to that produced by Professor Washington, barring a few translation differences (though that is an issue of the bilingual transducer). There are also some differences in the ordering of the grammatical glosses, since the autoglosser just uses the order of the tags according to the transducer.

04. Links

Access the autoglosser here:

https://github.com/Potato2357/autoglosser

05. Further Work

One place where further work may be done in the future is the ordering of the grammatical glosses. Perhaps some mechanism could be added to make sure the gloss is ordered in the most accurate way.
Of course, another place that needs work is to actually make the igt files for whatever language needs glossing.
Finally, another thing that might be useful is to make a tool to convert the text outputs of the autoglosser to something like LaTeX or MediaWiki.

06. Acknowledgements

I would like to thank Professor Jonathan Washington for teaching the Spring 2025 Computational Linguistics course and helping me with this project. I would also like to thank the Apertium Project for making this whole thing possible.