Морфологический анализатор языка руули с использованием формализмов *lexd* и *twol*

Филипп Бурлаков
научный руководитель О. А. Сеиков

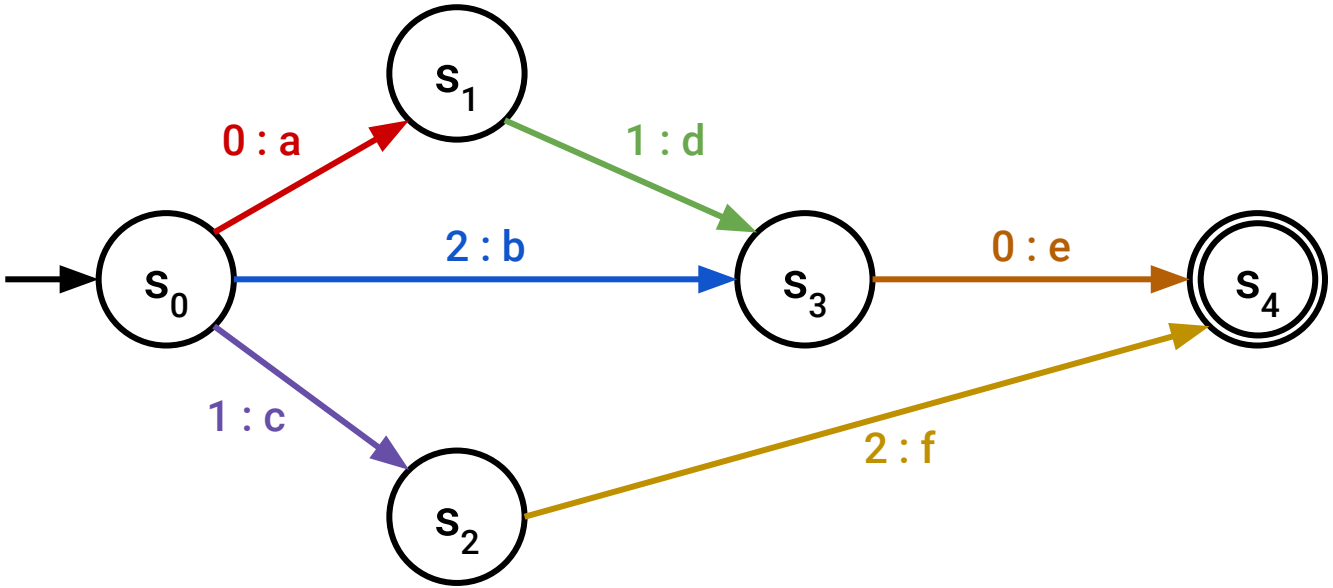
Введение. Общие понятия и инструменты

Морфологический анализатор – программа или алгоритм для извлечения морфологической информации словоформ.

HSFT формализмы (lexd и twol) – Helsinki Finite-State Transducer toolkit, предназначенный для создания морфологических анализаторов, включает в себя два формализма: lexd, предназначенный для описания морфотактики (глубинный уровень), и twol, описывающий фонологические, морфонологические и орфографические особенности (поверхностный уровень).

Формализм lexd представляет собой конечный автомат с выходом.

Введение. Конечный автомат с выходом



Вход	Выход
010 20 12	ade be cf

Введение. Информация о языке

Руули (известный также как рули, рууули-луньяла и др.) – малый язык группы банту, на котором говорят 160 000 человек в центральной Уганде.

Так как язык малоресурсный, создание морфологического анализатора с помощью методов глубинного обучения оказывается неэффективным, поэтому программа была создана на основе правил.

Этапы работы

1. Создание морфологического анализатора
    a. Изучение грамматики
    b. Написание правил
    c. Препроцессинг словаря

2. Оценка качества
    a. Препроцессинг данных
    b. Интерпретация результатов

3. Создание доступного интерфейса

Анализатор. Орфография

Алфавит выглядит следующим образом:

a, aa, b, bb, c, d, e, ee, f, g, i, ii, j, k, l, m, n, ny, ŋ, o, oo, p, r, s, t, u, uu, v, w, y, z

При этом для многих носителей *r* и *l* относятся к одной фонеме без определённой дистрибуции, поэтому было решено, что будет проще, если программа их тоже будет рассматривать как одну сущность.

Анализатор. Фонология и морфонология

- Фортиция глайдов
- Назальная ассимиляция
- Разрешение зияния
  - образование глайдов от *o*, *u*, *i*
  - прогрессивная ассимиляция гласного *a*
  - фузия гласного *a* последующим гласным (ограниченные контексты)
  - ? удаление гласного *a* или его переход в глайд *j* (ограниченные контексты, не описано в грамматике)

Остальные фонологические правила (например, палатализация) не влияют на орфографию, и поэтому не рассматривались.

Анализатор. Имена

Как и в других языках банту, в руули есть именные классы – словоклассифицирующая категория наподобие рода. У каждого класса свой префикс, и считается, что у каждого существительного есть “изначальный” класс.

Мена класса выражает регулярное значение, например множественность, и в правилах мы отделяем корень от префикса и прописываем его комбинаторные возможности, например:

(1) `<bank_account>:akawunti[9a,10a,naug]`

(2) `<mud>:nyangata[8,mass]`

Анализатор. Имена

При обработке словаря отсекается префикс “изначального” класса (который указан в словаре).

Проблема в том, что в словаре не учитываются периферийные “подклассы”, у которых префиксы и морфологическая сочетаемость отличаются от основного класса, но такое же согласование:

5	*e-*	*i-*	*(e)ibbaale* ‘stone’
5a	*e-*	*∅-*	*(e)dagala* ‘medicine’
5b	*e-*	*li- (ri-)*	*(e)riiso* ‘eye’

Анализатор. Глаголы

Глаголы оказались самой сложной частью речи по следующим причинам:

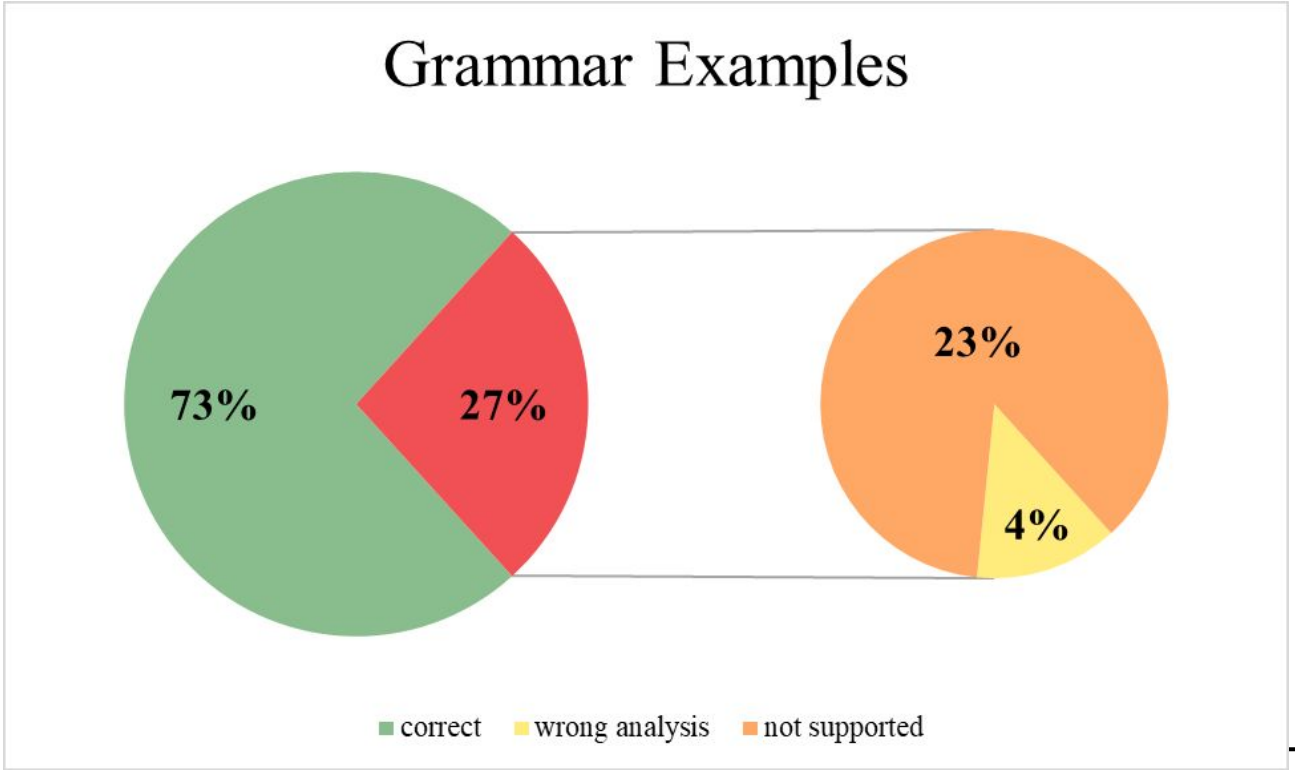
- сложные правила сочетания морфем между собой
- неописанные морфонологические процессы
- фузия перфективного суффикса с другими аффиксами (формы портманто)
- нерегулярные паттерны присоединения перфективного суффикса к корню

В итоге нерегулярные случаи присоединения суффикса префектива заносились в лексикон отдельно от простых глагольных корней.

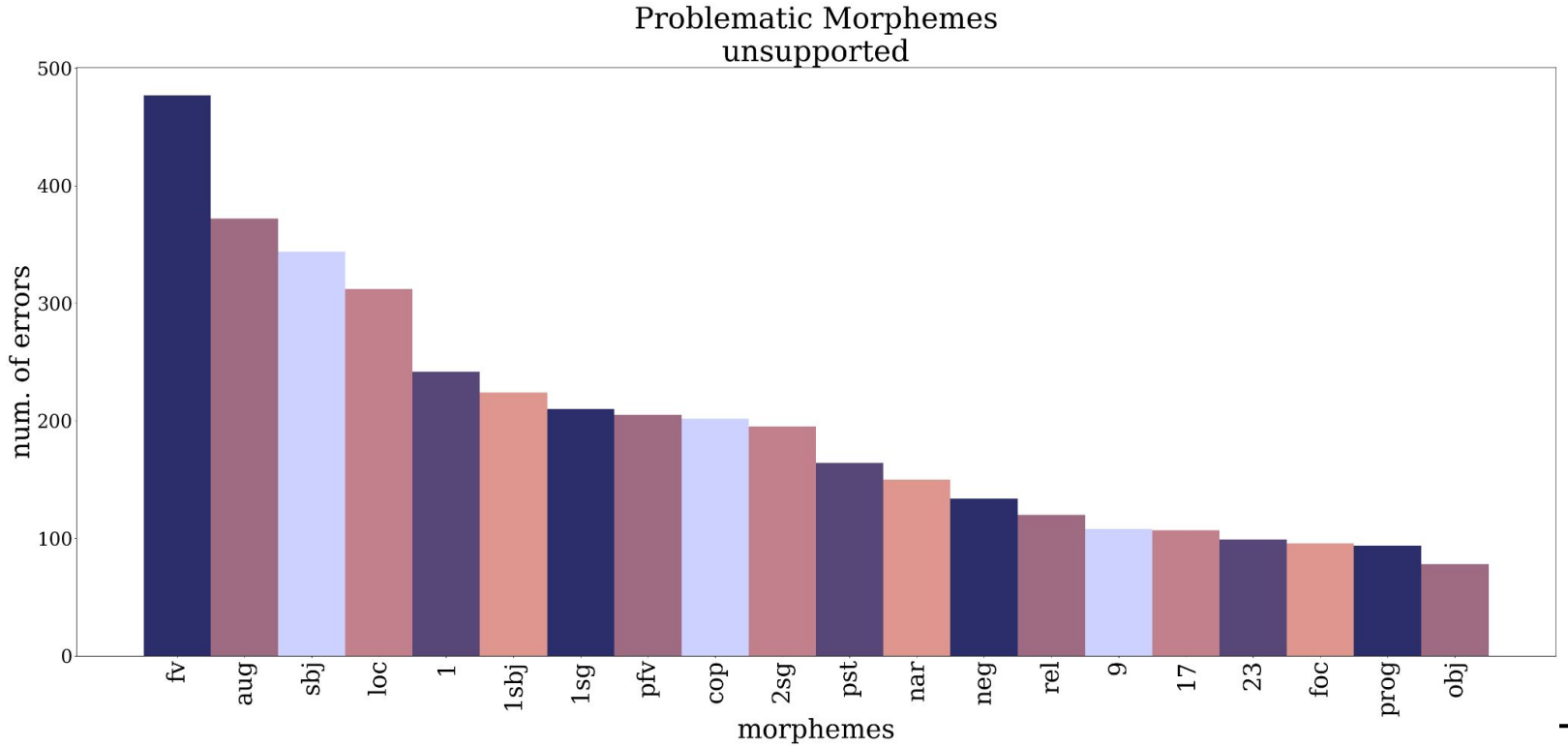
Оценка качества. Материал

корпус разобранных текстов в ELAN	примеры из грамматики
составлялся и обрабатывался автоматически	составлялся вручную, обрабатывался автоматически
большой (4896 словоформ)	маленький (113 словоформ)
много ошибок из-за отсутствия выравнивания разборов по словам в ELAN файлах	нет ошибок
разбор засчитывался, если в нём были представлены все глоссы из стандарта	разбор засчитывался при полном совпадении
подходит для выявления регулярных проблемных случаев	подходит для репрезентативной оценки качества

Оценка качества



Оценка качества. Проблемные случаи



Веб-интерфейс

Было разработано минималистичное веб-приложение, в котором можно подать на вход словоформу и получить на выходе варианты её глоссирования:

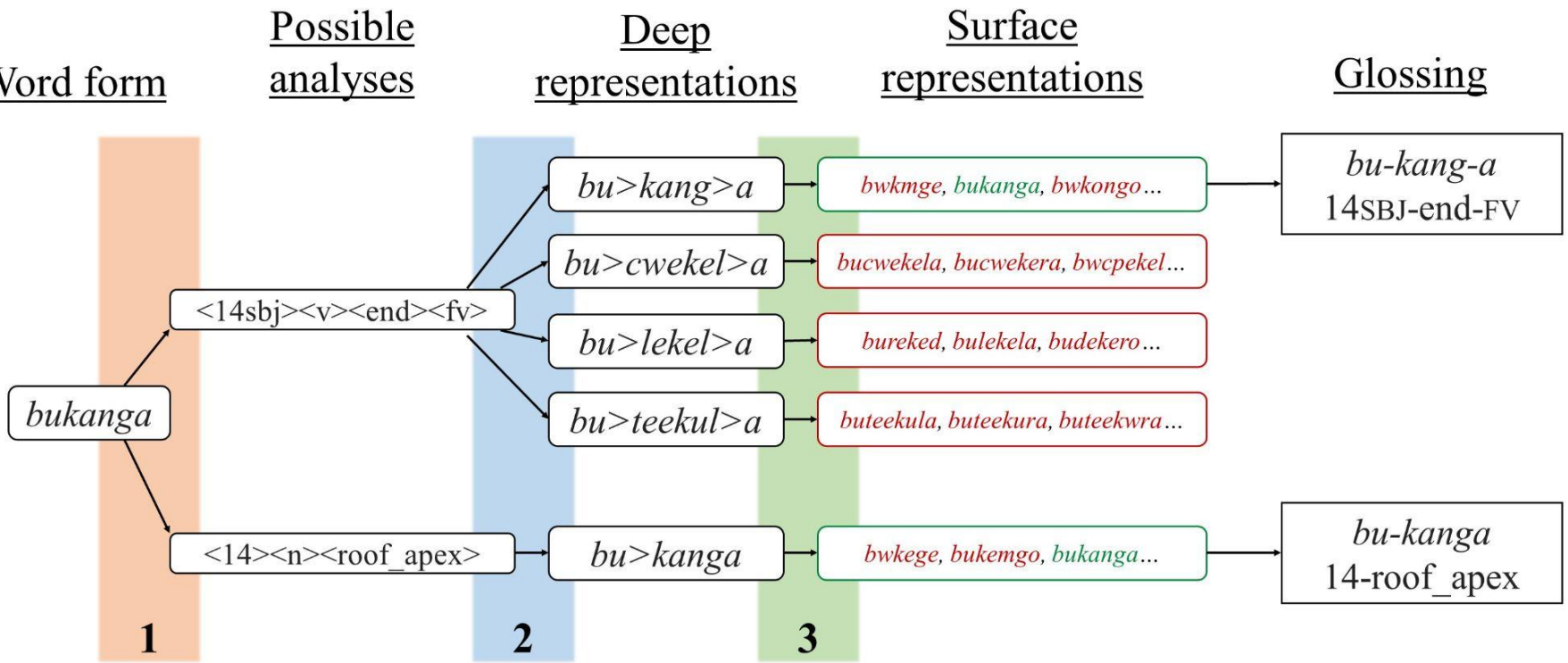
Веб-интерфейс

Было разработано минималистичное веб-приложение, в котором можно подать на вход словоформу и получить на выходе варианты её глоссирования:

Possible analyses of **tibakyakolesya:**

1. ti-ba-kya-kol-esy-a
   neg-2sbj-pers-do-caus-fv (v)
2. ti-ba-kya-kol-esy-a
   neg-2sbj-pers-handle-caus-fv (v)
3. ti-ba-kya-kolesy-a
   neg-2sbj-pers-use-fv (v)
4. ti-ba-kya-kol-esy-a
   neg-2sbj-pers-work-caus-fv (v)

Веб-интерфейс. Алгоритм



Заключение

- создана первая рабочая версия морфологического анализатора для языка руули, качество которой составляет 73%
- собраны случаи, с которыми возникают трудности
- создано удобное и простое приложение для получение глоссирования

В дальнейшем:

- работа над проблемными случаями
- доработка веб-интерфейса

Вопросы

Вопрос 1. В тексте указано, что перфективные формы доставались из словаря, что помогло избежать трудностей, связанных с нерегулярностью в их образовании. Есть ли уже какие-то идеи, как можно в дальнейшем работать с portmanteau forms, которые упоминались в работе?

Ответ. Сейчас суффиксы, задействованные в фузии этого типа рассматриваются как цельный элемент.

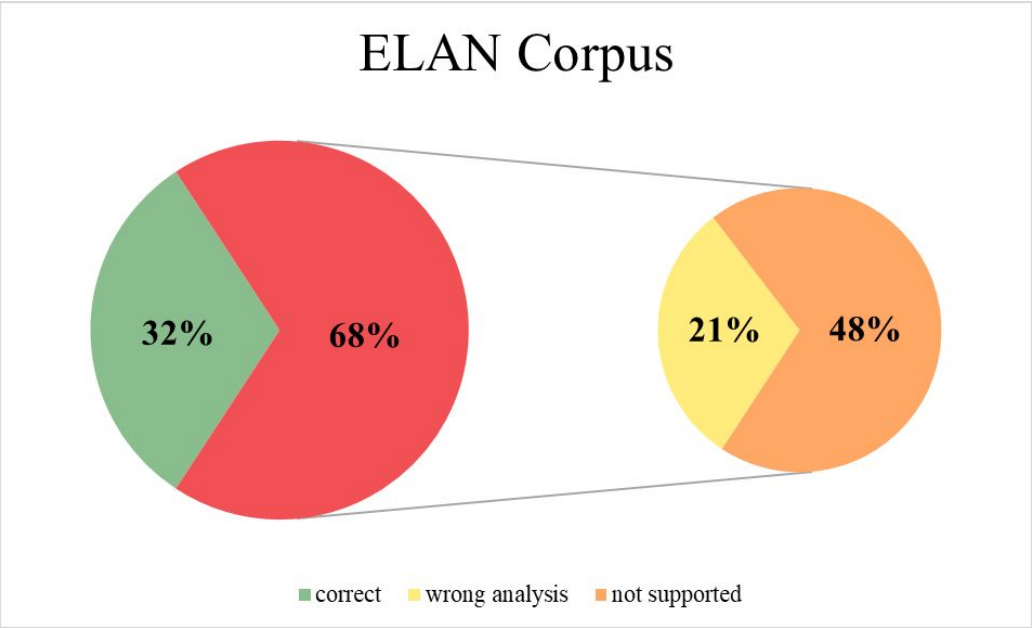
+ помогает избежать сложных морфонологических правил
- может не работать в случае, если дополнительно у суффикса перфектива происходит фузия с корнем.

Возможно, в будущем стоит рассматривать их в качестве отдельных элементов, но с дополнительными морфонологическими правилами

```
LEXICON PerfPortmanteau
<appl:pfv>:iile[i]
<caus:pfv>:isilye[i]
<caus:pfv>:ilye[i]
<appl:caus:pfv>:iilye[i]
<pass:pfv>:iibwe[i]
<appl:pfv>:eile[e]
<caus:pfv>:esilye[e]
<caus:pfv>:elye[e]
<appl:caus:pfv>:eilye[e]
<pass:pfv>:eibwe[e]
<recp:pfv>:angaine[i,e]
```

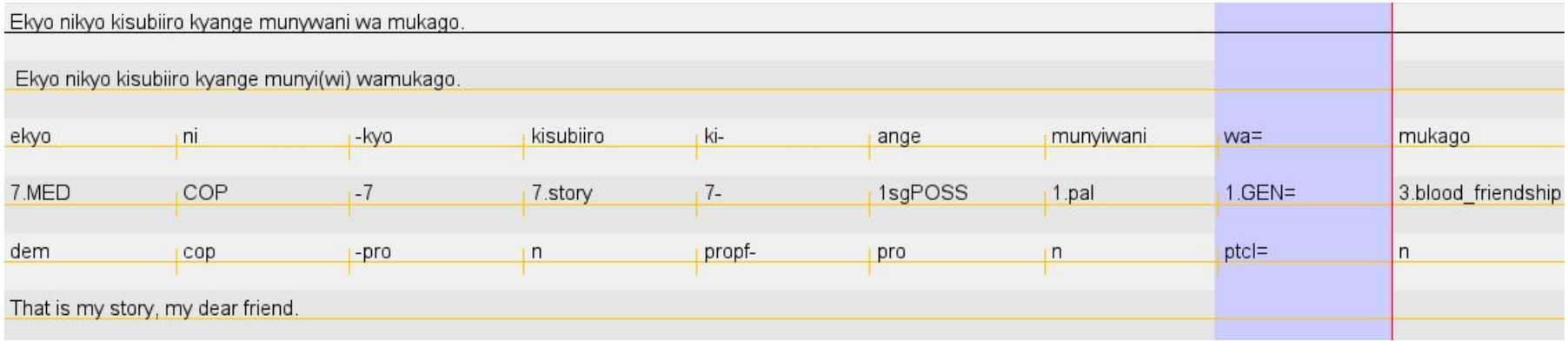
Вопросы

Вопрос 2. В заключении говорится, что точность работы анализатора может быть оценена на 73%. Этот вывод делается только исходя из результатов анализа второго корпуса? Кажется, из-за небольшого размера корпуса данная оценка может быть не совсем качественной. Может, стоит учитывать также первый корпус, но сделать вес его качества меньше, чем вес качества второго корпуса?

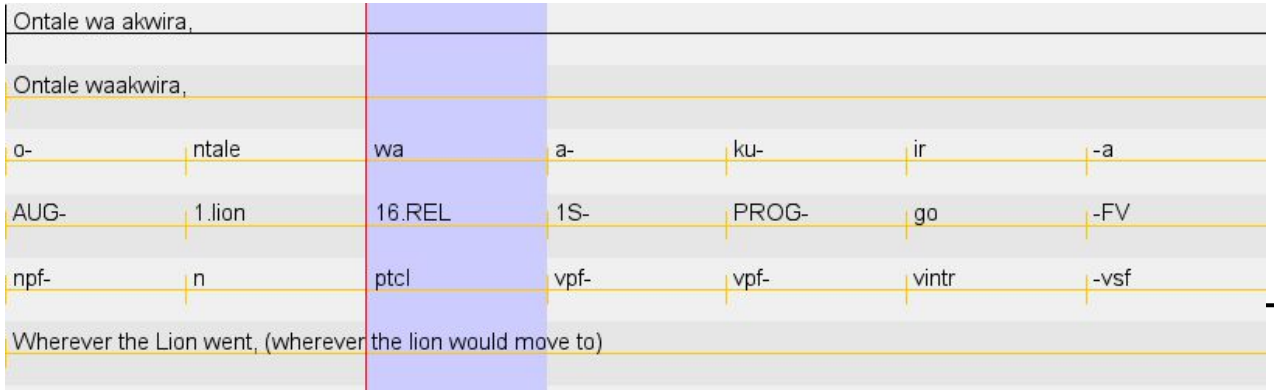


Вопросы

Вопрос 3. Можно ли каким-то образом до высчитывания процентов ошибок убрать из первого корпуса те примеры, где едет выравнивание?



Ekyo nikyo kisubiiro kyange munywani wa mukago.								
Ekyo nikyo kisubiiro kyange munyi(wi) wamukago.								
ekyo	ni	-kyo	kisubiiro	ki-	ange	munyiwani	wa=	mukago
7.MED	COP	-7	7.story	7-	1sgPOSS	1.pal	1.GEN=	3.blood_friendship
dem	cop	-pro	n	propf-	pro	n	ptcl=	n
That is my story, my dear friend.								



Ontale wa akwira,						
Ontale waakwira,						
o-	ntale	wa	a-	ku-	ir	-a
AUG-	1.lion	16.REL	1S-	PROG-	go	-FV
npf-	n	ptcl	vpf-	vpf-	vintr	-vsf
Wherever the Lion went, (wherever the lion would move to)						