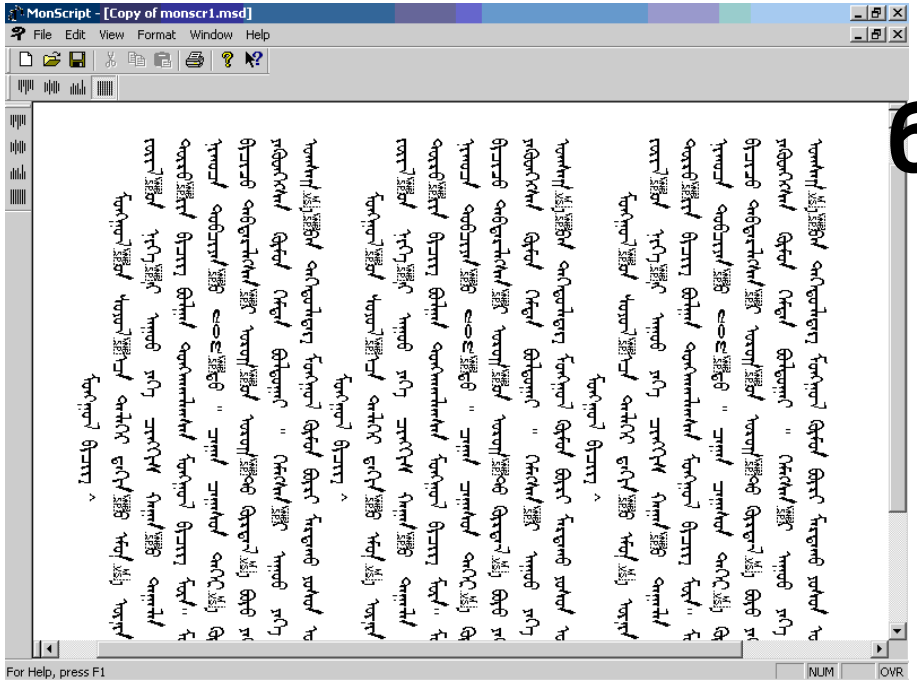

Текст боловсруулалтын үндэс

Лекц №4

Үгийн нормчлол ба үндсийг салгалт (stemming)

2020 он

Нормчлол

- “нормчлох” гэсэн ойлголт хэрэгтэй эсэх
  - Мэдээллийн хайлт: индекслэсэн текст болон хайлтын нэр томьёо ижил хэлбэртэй байх ёстой.
    - ***U.S.A.*** болон ***USA*** текстийг тэнцүүлэх хэрэгтэй
- Нэр томьёонуудын ижил байдлыг илээр тодорхойлдог
  - ж.нь., нэр томьёо доторх цэгийг устгах
- Хувилбар: тэгш хэмгүй тэлэлт:
  - Оролт: ***window*** Хайлт: ***window, windows***
  - Оролт: ***windows*** Хайлт: ***Windows, windows, window***
  - Оролт: ***Windows*** Хайлт: ***Windows***
- Алгоритмыг илүү хүчирхэг болсонч үр дүн бага байдаг

Том жижиг үсгийн асуудлыг багасгах нь

- Мэдээллийн хайлт(IR) шиг програмууд: бүх том үсгийг жижиг үсэг болгодог
  - Хэрэглэгчид жижиг үсгийг ашиглах хандлагатай байдаг
  - Алдаа гарах боломж: өгүүлбэрийн дундах том үсэг?
    - ж.нь., ***General Motors***
    - ***Fed*** vs. ***fed***
    - ***SAIL*** vs. ***sail***
- Хандлагийн шинжилгээний хувьд, Машин сургалт, Мэдээлэл гаргах
  - Том жижиг үсэг бас хэрэгтэй байдаг (***US*** болон ***us*** 2 том ялгаатай)

Леммачлал

- Үндсэн хэлбэрийн хувирал эсвэл хувилбарыг багасгах
  - *am, are, is* → *be*
  - *car, cars, car's, cars'* → *car*
- *the boy's cars are different colors* → *the boy car be different color*
- Леммачлал: толгой үгийн хэлбэрийг толиноос хайх
- Машин орчуулга
  - Spanish quiero (‘би хүсэх’), quieres (‘чи хүсэх’) same lemma as querer ‘хүсэх’

Үгийн хувилал

- **Бүтээвэрүүд**:
  - Үг бүтээх жижиг утгат нэгж
  - **Үндэс**: Цөм утгыг агуулсан нэгж
  - **Залгавар**: Үндсэн залгасан жижиг хэсэг
    - Дүрмийн үүрэгтэй байдаг
  - Залгаврууд

Үндсийг салгах - Stemming

- Мэдээллийн хайлтанд нэр томьёог үндэс болгож багасгадаг
- *Stemming* бол залгаврыг салгах
  - Хэлнээс хамааралтай
  - Жишээ нь, ***automate(s), automatic, automation*** бүгдийг ***automat*** болгоно.

for example compressed and compression are both accepted as equivalent to compress.

for exampl compress and compress ar both accept as equival to compress

Портерийн алгоритм
Англи хэлний хамгийн нийтлэг үндсийг салгагч

Алхам 1a

sses	→	ss	caresses	→	caress
ies	→	i	ponies	→	poni
ss	→	ss	caress	→	caress
s	→	ø	cats	→	cat

Алхам 1b

(*v*)ing	→	ø	walking	→	walk
			sing	→	sing
(*v*)ed	→	ø	plastered	→	plaster
…					

Алхам 2 (урт үндсийн хувьд)

ational	→	ate	relational	→	relate
izer	→	ize	digitizer	→	digitize
ator	→	ate	operator	→	operate
…					

Алхам 3 (урт үндсийн хувьд)

al	→	ø	revival	→	reviv
able	→	ø	adjustable	→	adjust
ate	→	ø	activate	→	activ
…					

Нэг корпус дахь үгийн хувирлыг харвал Яагаад ing –г хаяад байна? Хэрэв тэнд нэг эгшиг байвал?

$$(*v*)ing \rightarrow \emptyset$$

walking → walk

sing → sing

Нэг корпус дахь үгийн хувирлыг харвал Яагаад ing –г хаяад байна? Хэрэв тэнд нэг эгшиг байвал?

(*v*)ing → ø walking → walk
sing → sing

```
tr -sc 'A-Za-z' '\n' < shakes.txt | grep ’ing$' | sort | uniq -c | sort –nr
```

1312 King	548 being
548 being	541 nothing
541 nothing	152 something
388 king	145 coming
375 bring	130 morning
358 thing	122 having
307 ring	120 living
152 something	117 loving
145 coming	116 Being
130 morning	102 going

```
tr -sc 'A-Za-z' '\n' < shakes.txt | grep '[aeiou].*ing$' | sort | uniq -c | sort –nr
```

Нүсэр үгийн хувилал

- Зарим хэлүүд нүсэр бүтээврийн сегментлэл шаарддаг
  - Турк
  - Uygarlastiramadiklarimizdanmissinizcasina
  - `Хэрэв чи эдний дунд байгаад байвал бид хүмүүжиж чадахгүй’ гэж загнаж байгаа хэлбэр
  - Uygar `хүмүүжих’ + las `болох’
    + tir `шалтгаална’ + ama `боломжгүй’
    + dik `өнгөрсөн цаг’ + lar ‘олон тоо’
    + ...