ACTA MONGOLICA
INSTITUTE FOR MONGOLIAN STUDIES
National University of Mongolia

Vol.12 (383) 2012 172–177

ЭЛЕКТРОН БИЧВЭРТ СУУРИЛСАН МОНГОЛ ХЭЛНИЙ ЗӨВ БИЧИХҮЙН СУДАЛГАА

М.Энхжаргал, Ж.Пүрэв[1]

Түлхүүр үг

Амьд хэл боловсруулалт, корпус хэл шинжлэл, цахим хэл шинжлэл, зөв бичихзүй, монгол хэл

Товч агуулга

Мэдээллийн технологи болон бусад уулзвар салбар ухаанд үсрэнгүй хөгжиж буй чиглэлийн нэг нь амьд хэл боловсруулалт (Natural Language Processing) юм. Амьд хэл боловсруулалт гэдэг нь хэлийг машинаар боловсруулах судалгааны чиглэл бөгөөд компьютер, хэл шинжлэл, математик, электроник, хиймэл оюун ухаан, сэтгэл зүй зэрэг олон шинжлэх ухаанд хамаатай. Хэл боловсруулалтад зайлшгүй хэрэглэгдэх чухал зүйл бол тухайн хэлний талаарх мэдлэгийг бүрэн илэрхийлж чадахуйц үнэн бодит, их хэмжээний эхийн сан буюу материалын сан (corpus) юм.

Уг өгүүлэл нь МУИС-ийн Мэдээллийн технологийн сургуулийн Компьютер хэл шинжлэлийн судалгааны төвд (Center for research on language processing[2]) 2007-2009 он хүртэлх хоёр жилийн хугацаанд товлөрон хийгдсэн "Таван сая үгтэй монгол хэлний материалын сан"-д тулгуурлан электрон орчинд монгол хэлний зөв бичихүйн хэрэглээ, түвшин, соёлыг судалсан талаар өгүүлнэ.

Ерөнхий агуулга

1. Монгол хэлний материалын сан хэрхэн үүсгэсэн тухай

2. Материалын санд тулгуурлан компьютерийн орчинд (электрон орчинд) монгол хэлний зөв бичихүйн хэрэглээ.

1. Монгол хэлний материалын сан хэрхэн үүсгэсэн тухай

Хэл боловсруулалтад зайлшгүй хэрэглэгдэх чухал зүйл бол тухайн хэлний талаарх мэдлэгийг бүрэн илэрхийлж чадахуйц үнэн бодит, их хэмжээний эхийн сан буюу материалын сан (corpus) юм. Корпус нь дан ярианы, дан бичгийн эсвэл яриа ба бичгийн эх холилдсон их хэмжээний бичвэрийн сан бөгөөд корпус хэл шинжлэл нь цахим хэл шинжлэлийн нэг салбар

1. Монгол Улсын Их Сургууль
2. www.crlp.num.edu.mn

ухаанд тооцогддог. Корпус хэл шинжлэлийн ач холбогдол нь их бөгөөд олон талын судалгаанд хэрэгтэй. Ялангуяа хэл шинжлэлд цагаан толгойн үсгийн хэрэглэгдэх давтамж; үгсийн аймгийн өгүүлбэр дэх үүрэг, хэрэглээ; дагаврын чадамж, тархац; оносон хайлтаар мэдээлэл авах; төс утгатай үгийг ялгах; зөв бичих дүрмийн алдаа засах; толь бичгийн судалгаа гээд бүр цаашилбал машин орчуулгын судалгаанд үндэс суурь нь болж өгдөг.

Дэлхийд корпус бүтээх ажил 1960-аад оноос эхэлсэн бөгөөд орчин үеийн америкийн англи хэлийг судлах зорилгоор анх Brown корпусыг 1 сая орчим үгтэйгээр бүтээж байжээ. Түүнээс хойш улс орон бүр өөрийн хэлний үндэсний корпустай болох болсон бөгөөд LOB, London-Lund, AHI, British National Corpus, COBUILD зэрэг олон арван корпусууд зохиогджээ.

Хэлний корпуст тулгуурлан хэлний олон сонирхолтой дүгнэлтүүдийг хийж болно. Тухайлбал, Америкийн Brown корпус болон Британи англи хэлний LOB (Lancaster-Oslo-Bergen) корпусыг бүтээсний дараа эдгээр хоёр англи хэлний үгийн санг харьцуулсан судалгаа хийжээ. Энэ ажлаар зөв бичгийн colour/color, got/gotten сонирхолтой ялгааг илрүүлсэн байна [Ж.Пүрэв, 2006].

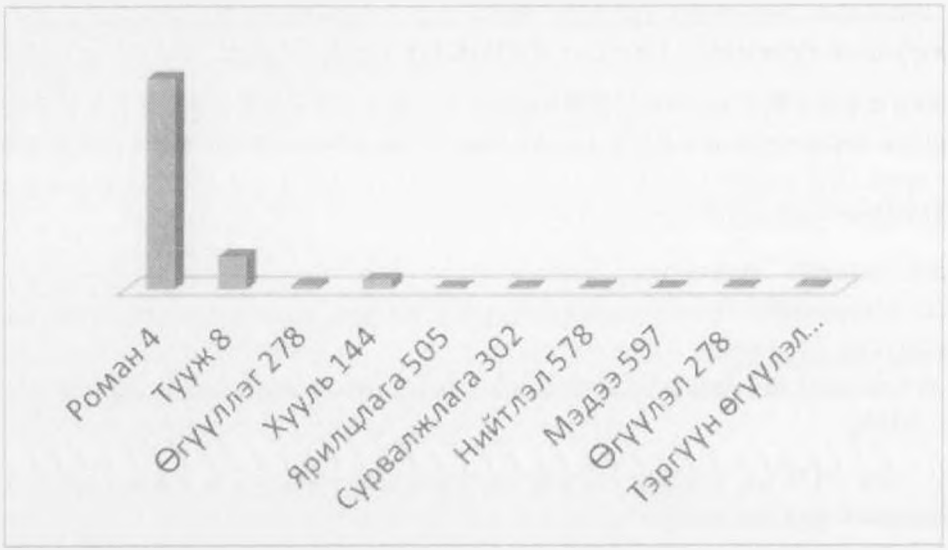
МУИС-ийн МТС-д Монгол хэлний материалын санг бүрдүүлэх ажил 2007-2009 оны хооронд Компьютер хэл шинжлэлийн судалгааны төвд төвлөрөн эхэлсэн бөгөөд одоогоор 5 сая орчим үгтэй материалын санг, холбогдох програмын системийн хамт бүрдүүлээд байна.

Монгол хэлний материалын санг цуглуулахад эхийг сонгохоос авахуулаад эх сурвалжуудыг хайж олох, эхийг эх сурвалжаас 'txt' хэлбэрт буулгах, эхийг цэвэрлэх, эхийг загварчлах, эхийн бүтцийг тодорхойлох зэрэг ажил багтана. Доор Зураг 1-т Монгол хэлний материалын санг найруулгын төрлөөр ангилан үзүүллээ.

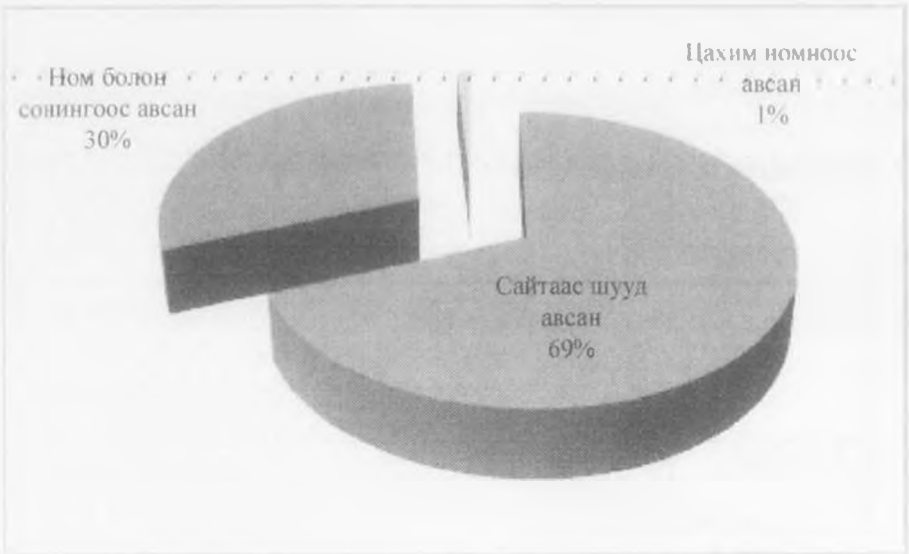
Найруулгын ангилал	Сайтаас авсан 2861 эх	Номноос авсан 24 эх	Сониноос авсан 1258 эх	Мултимедиа номноос авсан 10 эх
Хуулийн эх	www.legalinfo.mn			
Уран зохиолын эх	www.elibrary.mn www.mcl.edu.mn www.biirbeh.mn www.on-toli.mn	Л.Дашням Цаглашгүй цагаан салхи. Д.Пүрэвдорж Улаан зүүд.		Цэндийн Дамдинсүрэн. Монголын уран зохиолын дээж
Сонины эх	www.dailynews.mn www.zuuniimedee.mn		Үнэн сонин	

Монгол хэлний материалын сан нь нийт 5 сая орчим үгээс бүтэх бөгөөд давхардаагүй тоогоор 166000 орчим үг байна. Үүнд уран зохиолын эхээс роман, тууж, өгүүллэг; сонин нийтлэлийн эхээс мэдээ, тэргүүн өгүүлэл (Үнэн сонины), нийтлэл, ярилцлага, сурвалжлага, өгүүлэл; хуулийн эхээс үндсэн хууль болон бусад хуулийг сонгон авсан болно.

Зураг 1. Монгол хэлний материалын сангийн эх сурвалж



Зураг 2. Нэг эхэд буй үгийн дундаж тоо



Зураг 3.Материалын санг бүрдүүлсэн эх сурвалж

Эхийг цуглуулахад гар ажиллагааг хөнгөвчлөх зорилгоор 6 програмыг зохиосон. Үүнд, цахим хаягаас эхийг хуулах ажиллагааг нэг дор төвлөрүүлсэн 'TextCopier.exe' програм; эхийн зургийг авахад эвдэгдсэн, буруу үгийг засахын тулд 'Моозуур' нэртэй зөв бичгийн алдааг засах програм (Зураг 4); хавтсанд нэгтгэсэн 'doc' өргөтгөлтэй файлын үгийг тоолдог 'RCounter.exe' програм; хавтсанд нэгтгэсэн 'txt' өргөтгөлтэй файлын үгийг тоолдог 'TCounter.exe' програм; 'doc' өргөтгөлтэй файлыг 'txt' өргөтгөлтэй файл болгон хөрвүүлдэг 'TextConverter.exe' програм, төрөл бүрийн кодлолтой файлыг 'UTF-8' өргөтгөлтэй файл болгон хөрвүүлдэг 'Text file converter' програмыг тус тус зохиож хэрэглэсэн.

Зураг 4. Моозуур нэртэй зөв бичгийн алдааг засах програм

2. Материалын санд тулгуурлан компьютерийн орчинд (электрон орчинд) монгол хэлний зөв бичихүйн хэрэглээ.

Монгол хэлний материалын санд буй нийт үгийн 20 хувь нь алдаатай үг, 80 хувь нь алдаагүй байна. Энэхүү 20 хувьд багтаж буй алдааг төрөлжүүлэн үзвэл доорх байдлаар хуваах болох мэт санагдана. Үүнд,

a. Мэдлэгийн алдаа буюу зөв бичих дүрмийг зөрчиж бичсэн алдаа

- "ь"-ээр төгссөн үгэнд дагавар, нөхцөл залгахад "ь" нь "и"-ээр солигдох, эс солигдох дүрэм

- Жишээ нь: шуугидаг→шуугьдаг, халисан →хальсан.

- Эгшиг зохицох ёсыг зөрчиж болох и, ий, ы, эй эгшгүүдийг дагаж орох эгшиг нь тухайн эгшгийн өмнөх эгшигтэй зохицох дүрэм. Жишээ нь: төвөгтэйхэн→төвөгтэйхөн, төвийнхэнд→ төвийнхөнд.

- Үйл үгийн үндэс үүсгэх "–л" дагаварт шаардагдах балархай эгшгийг өмнө нь, хойно нь бичих дүрэм

- Жишээ нь: наслана→насална, цэгцлэж→цэгцэлж.

- Үйл үгийн үндэст нөхцөлдүүлэн холбох "–ж, –ч", "–жээ, –чээ" нөхцөлийг залгах дүрэм. Жишээ нь: өгөж→өгч, дарчээ→даржээ.

- Үйл үгийн үндэс үүсгэх "–р" дагаврыг залгах дүрэм. Жишээ нь: новшроно→ новширно.

- Балархай эгшиг гээгдэх дүрэм. Жишээ нь: мэхэлдэгийг→мэхэлдгийг, мэдээлсэнээр→ мэдээлснээр.

b. Хүн компьютерийн гар ашиглан бичих үед гарсан алдаа

- Компьютерийн гарын товчийн байрлал ойролцоо үгсийг сольсон

гялх →нялх, ёүхээ →сүхээ.

- Үсэг нэмсэн: ааав →аав, байгууууллагууд→байгууллагууд.

- Үсэг гээсэн: хрэгжүүлж →хэрэгжүүлж, гавшайлан →гавшгайлан.

- Үгийн үсгийн дарааллыг сольсон: банйа →байна, хийдэг →хийдэг.

- Кирилл, латин (роман) үсгийг хольж бичсэн: бүтээгдэхүүн →бүтээгдэхүүн, baримтаараа →баримтаараа.

- ASCII, Unicode хольж бичсэн: бітээгдэхіінээр →бүтээгдэхүүнээр, нэвтріілэх → нэвтрүүлэх.

- Бичлэг ижил үсэг цифрийг хольсон: 15ооо →15000, бөгөөд →бөгөөд.

c. OCR технологи буюу скайнердах үед гарсан алдаа

- Ойролцоо хэлбэртэй үсгийг сольсон

"ь" , "ъ" - гарья→гаръя, өгье→өгъе.

"ү", "у" - бурийн→бүрийн, дуурсэн→дүүрсэн.

- "т" , "г" - тоггнолыг→тогтнолыг, эмэггэйчүүд→эмэгтэйчүүд... гэх мэтээр ойролцоо хэлбэртэй нийт 28 үсгийг хооронд нь сольсон тохиолдол гарч байна.

- Тусдаа бичигдэх үгс нийлсэн: Барагхүн →бараг хүн, өдөралгасахгүй →өдөр алгасахгүй.

- Үсэг, цэг тэмдэг гээгдсэн: биелүүэхэд →биелүүлэхэд, үилдвэрлэл →үйлдвэрлэл

- Үсэг, цэг тэмдэг нэмэгдсэн: банримтаараа →баримтаараа, сарыӈ →сарын.

- Дундаа зураастай байх: бай-на →байна, бай-галь →байгаль.

НОМЗҮЙ

1. John Sinclair, Corpus, Concordance, Collocation. Oxford New York. 1996.

2. Насан-Урт С., Монгол хэл бичгийн сураг занги боловсруулах онол практикийн зарим асуудал, Улаанбаатар. 2004.

3. Graeme Kennedy. An introduction to corpus linguistics, London and New York. 1998.

4. Chagnaa Altangerel and Jaimai Purev: Web as a corpus. Mongol Studies Research papers, vol. XXXI (328). 2010. ISSN 1997-1826 (in Mongolian)

5. Равдан Э., Ж.Пүрэв. Тав дахь үеийн цахим хувьсгалаас үүдсэн монгол хэл соёлын тухай бодрол-ЭШБ. МУИС. ГХСС. №258/08/. УБ. 2006.

6. Боролзой Д., Пүрэвсүрэн Т. Компьютер хэл шинжлэл ба цахим үгийн сан байгуулах тухай асуудалд-Хэл зохиол судлал. ШУА, ХЗХ. Боть 1 (33). УБ., 2008.

7. Энхжаргал М., Таван сая үгтэй монгол хэлний материалын сан. МУИС, МТС, КХШСТ. 2008.

8. Алтангэрэл Г. Үгийн зөв бичих дүрмийн алдаа шалгуур програм. МУИС. МТС, 2008.

9. Пүрэв Ж. нар. Монгол хэлний тулгуур цахим байгууламж. МУИС. Азийн судлаагааны төв. Улаанбаатар. 2006.

RESUME

Electron text in internet is not only a wide range of text but behind it there is much knowledge and information. Therefore the information must be uploaded into the computer in correct and well arranged way from the first time.