

МОНГОЛ УЛСЫН ИХ СУРГУУЛЬ
ХЭРЭГЛЭЭНИЙ ШИНЖЛЭХ УХААН, ИНЖЕНЕРЧЛЭЛИЙН СУРГУУЛЬ

Ганболдын АМАРСАНАА

**ОЛОН ХЭЛНИЙ ЯЛГАМЖИТ ОНТОЛОГИЙГ
ОЛНЫ ХҮЧЭЭР НУТАГШУУЛАХ АРГАЧЛАЛ**
**(A Methodology for Diversity-aware Multilingual
Ontology Localization via Crowdsourcing)**

Мэргэжил: F524300, Мэдээллийн технологи

Улаанбаатар, 2016

МОНГОЛ УЛСЫН ИХ СУРГУУЛЬ
ХЭРЭГЛЭЭНИЙ ШИНЖЛЭХ УХААН, ИНЖЕНЕРЧЛЭЛИЙН СУРГУУЛЬ

Ганболдын АМАРСАНАА

**ОЛОН ХЭЛНИЙ ЯЛГАМЖИТ ОНТОЛОГИЙГ
ОЛНЫ ХҮЧЭЭР НУТАГШУУЛАХ АРГАЧЛАЛ**

**(A Methodology for Diversity-aware Multilingual
Ontology Localization via Crowdsourcing)**

Доктор (PhD)-ын зэрэг горилсон нэг сэдэвт бүтээл

Эрдэм шинжилгээний ажлын

Удирдагч: _____ Доктор (PhD), дэд профессор Ч.Алтангэрэл

Хамтран удирдагч: _____ Доктор (PhD), профессор П.Жүнхилья

Зөвлөх: _____ Доктор (PhD), профессор А.Эрдэнэбаатар

_____ Доктор (PhD), дэд профессор Ч.Лодойравсал

Улаанбаатар, 2016

Гарчиг

Оршил.....	1
БҮЛЭГ 1. Онтологи ба ялгамжийн асуудал	5
1.1 Мэдлэгийн сан ба бодит ертөнцийн загварчлал	5
1.2 Ертөнцийн мэдлэгийн сан	10
1.3 Олон хэлний онтологи нутагшуулалт	15
1.4 Ялгамжийн асуудал	24
Бүлгийн дүгнэлт.....	27
БҮЛЭГ 2. Олны хүчийг ашиглах судалгаа	29
2.1 Олны хүч.....	29
2.2 Орчуулга ба олны хүч.....	35
2.3 Онтологи нутагшуулахад олны хүчийг ашиглах боломж	42
Бүлгийн дүгнэлт.....	50
БҮЛЭГ 3. Олны хүчээр онтологи нутагшуулах аргачлал	51
3.1 Онтологи нутагшуулах ажил	51
3.2 Мэргэжилтнүүдээр онтологи нутагшуулах аргачлал.....	59
3.3 Вэб хэрэглэгчдээр онтологи нутагшуулах аргачлал	83
3.4 Үр дүн	103
Дүгнэлт	109
Ном зүй.....	111
Хавсралт А. Ертөнцийн мэдлэгийн сангийн холбоос	117
Хавсралт Б. Файлын хүснэгтэн загварын жишээ	118
Хавсралт В. Хэлний цөмийг экспортлох RDF файлын жишээ	120
Хавсралт Г. НМЦ-ийг нутагшуулах системийн хэрэглэгчийн зарим интерфейс	123
Хавсралт Д. Тохирлын үзүүлэлтүүдийг тооцоолох R скрипт	127
Хавсралт Е. Дүйцэлгүй ойлголтын жишээ.....	129

Товчилсон үгийн жагсаалт

UKC	– Universal Knowledge Core
LKC	– Local Knowledge Core
IoT	– Internet of Things
RDF	– Resource Description Framework
W3C	– World Wide Web Consortium
URL	– Uniform Resource Locator
ABox	– Assertion Box
TBox	– Terminological Box
DERA	– Domain-Entity-Relation-Attribute
FBK	– Fondazione Bruno Kessler
ERD	– Entity Relationship Diagram
SKOS	– Simple Knowledge Organization System
UML	– Unified Modelling Language
DL	– Description Logics
CG	– Conceptual Graphs
FOL	– First Order Logic
OWL	– Web Ontology Language
ISO	– International Standard Organization
XBRL	– eXtensible Business Reporting Language
RDFS	– Resource Description Framework Schema
YAGO	– Yet Another Great Ontology
UI	– User Interface
CIA	– Central Intelligence Agency
ISO	– International Organization for Standardization
LIR	– Linguistic Information Repository
HIT	– Human Intelligence Task
ESP	– Extra Sensory Perception
OCR	– Optical Character Recognition
BLUE	– Bilingual Evaluation Understudy

HTER	– Human-mediated Translation Edit Rate
WER	– Word Error Rate
TER	– Translation Edit Rate
DARPA	– Defense Advanced Research Projects Agency
GALE	– Global Autonomous Language Exploitation
WMT	– Workshop on Statistical Machine Translation
ICC	– Inter-Class Correlation
UB	– User Base
CSV	– Comma Separated Values
MTurk	– Amazon Mechanical Turk
NT	– Narrower Term
BT	– Broader Term
MKS	– Meter Kilogram Second
NLCore	– Natural Language Core
GAP	– Lexical Gap
UK ID	– Universal Knowledge Identifier
WSD	– Word Sense Disambiguation
UPM	– Universidad Politécnica de Madrid
NUM	– National University of Mongolia
UNITN	– University of Trento
NUIG	– National University of Ireland Galway
WWW	– World Wide Web
ЕМС	– Ертөнцийн мэдлэгийн сан
НМЦ	– Нутгийн мэдлэгийн цөм
НОС	– Нэгж-объектийн сан
ХОД	– Хүний оюуны даалгавар

Нэр томъёоны тайлбар

нэгж-объект	(<i>entity</i>) бодит ертөнц дээр оршин буй биет болон хийсвэр, объектив шинж чанартай ямар нэг зүйл;
өгөгдөл	(<i>data</i>) аливаа үйл явдал, баримтыг илэрхийлсэн, тодорхой бүтэцтэй бичвэр эсвэл тоон утгын цуглуулга;
мета-өгөгдөл	(<i>meta-data</i>) мэдээлэл эсвэл өгөгдлийн бүрдлийн бүтэц, формат, чанар, агуулга, бэлтгэн нийлүүлэгч, эх үүсвэр, боловсруулсан арга зүй, байрлалыг тодорхойлсон эсвэл тайлбарласан өгөгдөл;
мэдээлэл	(<i>information</i>) аливаа харилцаа эсвэл өгөгдөл, баримтын талаарх мэдлэгийг илэрхийлж байгаа бичвэр, тоо, график, дуу, дүрс эсвэл газрын зураг;
мэдлэг	(<i>knowledge</i>) туршлага, эсвэл боловсролд тулгуурлан олж авсан цэгцэлсэн мэдээлэл;
мэдлэгийн сан	(<i>knowledge base</i>) аливаа зүйлсийн төрөл, шинж чанар, тэдний хоорондын холбоо хамаарлыг, тэдгээр зүйлсийн бодит тохиолдлууд болох нэгж-объектууд, тэдгээрийн шинж, хоорондын холбоо хамаарлыг хадгалдаг цахим сан;
онтологи	(<i>ontology</i>) хоорондоо утгазүйн холбоос бүхий ойлголтуудын олонлог;
ай	(<i>domain</i>) аливаа харилцаанд хэрэглэх ямар нэг мэдлэгийн салбар эсвэл судлах сэдэв;
фасет	(<i>facet</i>) аливаа мэдлэгийг олон талаас нь тодорхойлох төрөл ойлголтуудын шатлал;
үглэвэр	(<i>lemma</i>) үгийн гадаад хэлбэр буюу үгийн бичигдэх, дуудагдах байдал;
утгалбар	(<i>sense</i>) үгийн агуулга, утга;
синсет	(<i>synset</i>) ойролцоо утгаар хэрэглэж болдог үгсийн олонлог;
дүйцэлгүй ойлголт	(<i>concept gap</i>) ямар нэг хэл соёлд хэрэглэдэггүй ойлголт;
ялгамж	(<i>diversity</i>) аливаа зүйлсийн ялгаатай байдал эсвэл ялгаатай үзэгдэл;
олны хүч	(<i>crowdsourcing</i>) олон нийтийн нэг хүн бүрийн хувь нэмрийг их хэмжээгээр цуглуулж хүний оюуны ашгийг хүртэх үйл явц;
хүний оюуны даалгавар	(<i>human intelligence task</i>) компьютер гүйцэтгэхэд хэцүү, хүмүүст хялбар даалгавар;

Хүснэгтийн жагсаалт

Хүснэгт 1.1 Түгээмэл мэдлэгийн сангуудын агуулгын тоон харьцуулалт [40].....	9
Хүснэгт 1.2 Онтологи нутагшуулалтын төрөл [16].....	22
Хүснэгт 2.1 Олны хүч ашигласан системийн төрлүүд [51].....	30
Хүснэгт 2.2 Оролцогчдыг авч үлдэх арга.....	32
Хүснэгт 2.3 Статистикийн хэмжүүрүүдийн ангилал	45
Хүснэгт 2.4 Үнэн өгөгдлийн матриц	49
Хүснэгт 2.5 Элементүүд доторх тохиролцлын хүснэгт	49
Хүснэгт 2.6 Нэрийдсэн ангиллын зөрүүгийн функц.....	50
Хүснэгт 3.1 Нутагшуулах онтологийн элементүүд.....	51
Хүснэгт 3.2 Олны хүчээр хийх ажлууд	57
Хүснэгт 3.3 Нутгийн мэдлэгийн цөмийн мэргэжилтнүүдийн үүрэг	67
Хүснэгт 3.4 RDF болон файлын хүснэгтэн загвар хоорондох буулгалтын жишээ	78
Хүснэгт 3.5 Орон зайн онтологийг нутагшуулсан үр дүн	82
Хүснэгт 3.6 Синсетийн орчуулгад санал болгосон үгс.....	90
Хүснэгт 3.7 Синсетийн орчуулгад санал өгсөн байдал	91
Хүснэгт 3.8 Синсетэд санал өгсөн үнэн өгөгдлийн матриц	92
Хүснэгт 3.9 Синсетийн үгсийн ангиллын тохиролцлын хүснэгт	92
Хүснэгт 3.10 Синсетэд санал өгсөн үнэн өгөгдлийн матриц (орхисон өгөгдөлтэй)... ..	93
Хүснэгт 3.11 Синсетийн үгсийн ангиллын тохиролцлын хүснэгт (орхисон өгөгдөлтэй).....	93
Хүснэгт 3.12 Хүний оюуны даалгаврын гүйцэтгэлийн тоон үзүүлэлт	94
Хүснэгт 3.13 Жишиг сангийн хэмжээ	94
Хүснэгт 3.14 Коэффициентын утга дахь тохирлын үзүүлэлтүүд	101
Хүснэгт 3.15 Үр дүнгийн үзүүлэлтүүдийн харьцуулалт	106
Хүснэгт 3.16 Онтологи нутагшуулах аргачлалын ерөнхий харьцуулалт	107
Хүснэгт 3.17 Олны хүчээр онтологи нутагшуулах аргачлалын туршилт, үр дүнгийн харьцуулалт	108

Зургийн жагсаалт

Зураг 1.1 Өгөгдлөөс мэдлэг үүсэх шатлал	6
Зураг 1.2 Мэдлэгийн сангийн жишээ бүдүүвч.....	8
Зураг 1.3 Гурвалын график дүрслэл	8
Зураг 1.4 RDF жишээ.....	9
Зураг 1.5 Ертөнцийн мэдлэгийн сангийн бүтэц	11
Зураг 1.6 Хэлний цөмийн нэгж-объект холбоосны (ERD) диаграм	14
Зураг 1.7 Хэл болон формаль талаас ангилсан онтологийн төрөл [11].....	15
Зураг 1.8 Айн зорилгод түшиглэсэн онтологийн ангилал [11]	16
Зураг 1.9 Ангилал онтологийн жишээ [12]	17
Зураг 1.10 Ертөнцийн мэдлэгийн сангийн ялгамж.....	25
Зураг 2.1 Олны хүч ба бусад судалгааны чиглэл [49].....	29
Зураг 2.2 Хамтын ажиллагаат орчуулгын нэг загвар	36
Зураг 2.3 Олны хүчээр гүйцэтгэсэн орчуулгын BLEU оноог олон аргаар бодсон туршилтын үр дүн [54].....	39
Зураг 3.1 Онтологи нутагшуулах ажлын ерөнхий даалгавар.....	52
Зураг 3.2 Сацрал үгсээр илэрхийлэх ойлголт	54
Зураг 3.3 Дүйцэлгүй ойлголтын хувилбар	54
Зураг 3.4 Үгийн утгалбарын ялгамж	55
Зураг 3.5 Нутгийн мэдлэгийн цөм ба Ертөнцийн мэдлэгийн сангийн уялдаа	60
Зураг 3.6 Нутгийн мэдлэгийн цөмд синсетийг орчуулах макро алхамууд.....	60
Зураг 3.7 Орон зайн айн газарзүйн тогтоц фасетийн дэд хэсэг	63
Зураг 3.8 Ойлголтыг орчуулах үйл явцын диаграм	65
Зураг 3.9 Шинэ ойлголт нэмэх үйл явцын диаграм	66
Зураг 3.10 Гарал үүслийн UML диаграм.....	69
Зураг 3.11 НМЦ системийн архитектур	71
Зураг 3.12 Орчуулах болон үнэлэх үйл явцын дарааллын диаграм	73
Зураг 3.13 LKC-UI-ийн НМЦ орчуулагч хэрэглэгчийн интерфейс.....	74
Зураг 3.14 LKC-UI-ийн гарал үүслийн бүртгэлийн цонх	75
Зураг 3.15 Хүснэгтэн загвараар хөгжүүлсэн өгөгдлийг RDF-д буулгасан жишээ.....	79
Зураг 3.16 Орон зайн айн фасетуудын дэд хэсэг	80
Зураг 3.17 Орон зайн холбоос (<i>R</i>) фасетийн дэд хэсэг	80
Зураг 3.18 хур шинжийн (<i>A</i>) фасетийн дэд хэсэг.....	81
Зураг 3.19 Вэб хэрэглэгчдээр онтологи нутагшуулах аргачлалын бүдүүвч.....	85

Зураг 3.20 Синсет орчуулах даалгаврын хэрэглэгчийн интерфейс	87
Зураг 3.21 Синсет үнэлэх даалгаврын хэрэглэгчийн интерфейс	89
Зураг 3.22 Санал нийцлийн хэмжүүрийн утгаас хамаарсан синсетийн давтамж	95
Зураг 3.23 Синсетийн үгсийн тооноос хамаарсан санал нийцэл	96
Зураг 3.24 Санал нийцлийн коэффициентын утгууд дээрх синсетийн тохирлын үзүүлэлтүүд	97
Зураг 3.25 Санал нийцлийн хэмжүүрүүдийн тохирлын үзүүлэлтүүд	99
Зураг 3.26 Тохирлын үзүүлэлтүүдийн хоорондын харьцуулалт	100
Зураг 3.27 Вэб хэрэглэгчдийн гүйцэтгэсэн ажлын чанар	101
Зураг 3.28 Дийлэнх олонхын саналаар сонгосон үгсийг орчуулсан болон үнэлсэн вэб хэрэглэгчдийн ажлын чанар	102
Зураг 3.29 НМЦ-ЕМС систем бүрэлдэхүүний диаграм	104
Зураг 3.30 Олны хүчээр онтологи нутагшуулах аргачлалын алгоритм	105

ОРШИЛ

Хүний хүсэл хэрэгцээг эх хэлээр нь ойлгож оновчтой үйлдэл хийж чаддаг илүү ухаалаг онлайн үйлчилгээ ирээдүйн вэб технологи – семантик вэбийн [1]–[3] хөгжлийг тодорхойлж байна. Энэ технологид хүмүүсийн бодит үйл явдал, баримтыг үлэмж хэмжээгээр цэгцлэн хадгалсан мэдлэгийн сан [4]–[6] шаардлагатай. Ийм сан орчлон ертөнцийн ойлголтуудыг (*уул, нуруу* г.м.) ангилж тэдгээрийн хоорондын уялдаа холбоог (*тэр бол ..., ... бол бүрдэл* г.м.) загварчилсан онтологийг агуулдаг. Жишээ нь, *уул бол өндөрлөг газар, өндөрлөг газар бол геологийн тогтоц* гэх мэт. Хэдэн зуун мянган ойлголтыг олон хэлээр илэрхийлэх онтологиуд хэрэглээнд нэвтэрч эхэлсэн боловч хэл соёлын ялгааг үгийн утгын, ойлголтын түвшинд нарийн тусгасан, соёл хооронд чөлөөтэй ашиглах нэгдмэл онтологийг үүсгэх асуудал бүрэн шийдэгдээгүй байна. Энэ нь даярчлагдаж буй цахим ертөнцөд олон эх сурвалж, хэл соёлоос мэдээллийг боловсруулж мэдлэг, туршлага солилцон хамтдаа хөгжих үндэстэн дамнасан харилцаанд маш чухал билээ.

Онтологийг үүсгэхэд мэргэжлийн хүч, цаг хугацаа их шаарддаг. Тэр дундаа цахим нөөц багатай хэл соёлын (жишээ нь, Монгол) хувьд өмнө хөгжүүлсэн ямар нэг онтологи – эх онтологийг нутагшуулах гар аргаар [7] шинэ онтологи үүсгэхэд эдгээр асуудлыг тодорхой хэмжээнд шийдвэрлэх, эх онтологийн бүтцийн ихэнх хэсгийг шууд авч болдог давуу талтай. Онтологийг гар аргаар [8], [9] үүсгэх нь илүү их мэргэжлийн хүч зарцуулдаг ч өндөр чанартай байдаг. Харин хагас автомат аргаар үүсгэхэд олон хэлний тайлбар толь бичиг, үгийн сангийн өгөгдлийн сан, бичвэрийн материалын сан зэрэг хэлний нөөц, үгийн утга салгах, машин орчуулга зэрэг програм хангамжийг шаарддаг [7], [10]. Үүнийг хязгаарлагдмал нөөцтэй, хэл боловсруулалтын технологи сайн хөгжөөгүй хэл соёлын хувьд хэрэгжүүлэх боломжгүй байдаг.

Иймд олон хэл соёл хоорондын онцлог шинж, ялгаатай байдлыг (ялгамж) чанарын өндөр түвшинд тусгасан, нэгдмэл онтологийг богино хугацаанд бага зардлаар, хэлний нөөцөөс хамааралгүй үүсгэх боломжийг судалж шинэ аргачлал боловсруулах шаардлагатай байна.

Судалгааны ажлын зорилго, зорилт. Энэ ажлаар олон хэл соёлын ялгамжийг хэлний болон ойлголтын хүрээнд тусгасан онтологийг олны хүч ашиглан нутагшуулах аргачлалыг боловсруулах, турших, үнэлэх зорилготой. Үүнд дараах зорилтуудыг дэвшүүлэв.

1. Онтологийг нутагшуулахад олны хүчийг ашиглах боломжийг судлах
2. Ялгамжийн төрлийг тодорхойлох
3. Ялгамжийг тусгасан онтологийг нутагшуулах аргачлалыг боловсруулах
4. Боловсруулсан аргачлалыг туршиж, үнэлэх

Судлах зүйл. Дэвшүүлсэн зорилтууддаа хүрэхийн тулд олон хэлний онтологи, онтологийг нутагшуулах үйл явц, хэл соёл хоорондох ойлголт болон түүнийг илэрхийлэх үгийн сангийн ялгамж, олны хүч, олны хүчээр гүйцэтгэсэн ажлын чанарын үнэлгээ зэргийг судлах юм.

Судалгааны арга. Анализ, туршилт, харьцуулан судлах зэрэг судалгааны нийтлэг аргуудаар ялгамжийн төрөл, ялгаатай ойлголтуудыг илэрхийлэх үгийн сангийн нэгж, онтологи нутагшуулах үйл явц, олны хүчийг судалсан. Онтологи нутагшуулах аргачлалыг туршилтын аргаар, олны гүйцэтгэсэн ажлыг нэгтгэх үйл явцыг асуумжийн аргаар, чанарын үнэлгээг тоон баримт цуглуулах болон индукцийн аргуудыг хослуулан судалсан.

Сэдвийн судлагдсан байдал. Олон улсын түвшинд К.Рөүсси нар [11] онтологийн үндсэн төрөл зүйлийг тодорхойлсон, П.Жүнхилья нар [12]–[14] онтологийг хэлбэр, ойлголтын холбоосны ерөнхий төрөл талаас хоёр ангилсан байдаг. Онтологийг нутагшуулах автомат болон хагас автомат аргуудыг судлаач М.Эспиноза нар [7], [10], [15], [16], М.Аркан нар [17], [18], С.Валтер нар [19], онтологийн үгийн сангийн загварыг П.Чимиано нар [20], цахим нөөц багатай хэлний толь бичгийг өгөгдлийн олборлолтоор гаргах тухай М.Бенжамин нар [21], гар аргаар үгийн сангийн өгөгдлийн сан үүсгэх асуудлыг Ж.Миллер нар [8], үүнээс гар орчуулгын аргаар шинэ өгөгдлийн сан үүсгэх ажлыг К.Линден [9] нар тус тус судалсан байна. Онтологийн үгийн санг олны хүчээр үүсгэх аргачлалыг Б.Лансер нар [22] судалсан. Монгол хэлэнд ВөрдНэтийг автоматаар үүсгэх арга болон ВөрдНэтийн менежмент системийн судалгааг Хаси нар [23] хийсэн байна. Ч.Алтангэрэл нар [24] Монгол хэлний үгийн утгазүйн сүлжээг толь бичиг ашиглан автоматаар үүсгэх ажлыг гүйцэтгэжээ.

Судалгааны ажлын таамаглал. Олон хэл соёлын нарийн зааг ялгааг тусгасан онтологийг олны хүчийг оновчтой зохион байгуулснаар өндөр чанартайгаар, богино хугацаанд үүсгэх боломжтой гэж үзсэн. Олон нийтийн оруулж буй хувь нэмрийг нэгтгэхэдээ оролцогчдын дотоод санал нийцэл өндөр байхад үнэн үр дүн гарган авч болно.

Судалгааны ажлын шинэлэг тал. Онтологийг нутагшуулахад хязгаарлагдмал нөөцтэй хэл соёлын хувьд бусдын хөгжүүлсэн сайн нөөцийг авч ашиглахын зэрэгцээ олны хүчээр чанартай өгөгдөл гарган авах замаар богино хугацаанд ялгамжийг тусгасан нэгдмэл онтологийг үүсгэж болохыг харуулсан. Түүнээс гадна олны хүчний оруулсан хувь нэмрийг үнэлэхэд хүмүүсийн дотоод санал нийцлийг тооцох статистик хэмжүүрийг онтологи нутагшуулах ажилд шинээр нэвтрүүлсэн болно.

Практик ач холбогдол. Боловсруулсан аргачлалын туршилтын үр дүнд орон зайн айн 1,042 ойлголт бүхий монгол онтологийг үүсгэж практик дээр харуулж чадлаа. Энэ аргачлалын дагуу Италийн Трэнтогийн их сургуулийн KnowDive судалгааны групп програм хангамжийн систем хөгжүүлсэн бөгөөд үүнийг Монгол, Мандарин, Испани, Итали, Амхарик гэх зэрэг 10 гаруй хэл дээр амжилттай хэрэгжүүлж байна.

Хамгаалахаар дэвшүүлж буй асуудлууд. Энэ ажлаар дараах дэд асуудлуудыг хамгаалахаар дэвшүүлж байна.

1. Ялгамжит онтологийг мэргэжлийн олны хүчээр нутагшуулах аргачлал
2. Вэб хэрэглэгчдээр онтологи нутагшуулах арга
3. Олны оруулсан хувь нэмрийг Криппендорфийн альфа, Флайсын каппа статистик хэмжүүрээр үнэлэх арга
4. Олны хүчээр нутагшуулсан онтологийн чанар, түүний үнэлгээ

Үр дүнг хэлэлцүүлсэн байдал. Энэ судалгааны үр дүнгүүдийг

1. мэргэжлийн түвшинд хянан магадалгаа хийгддэг, Scopus SJR индекстэй, олон улсын мэргэжлийн сэтгүүлд 1 өгүүлэл [25],
2. эрдэм шинжилгээний олон улсын хурлын эмхтгэлд (Итали, Энэтхэг, Испани, Хятад улсад) нийт 4 [26]–[29] өгүүлэл тус тус хэвлүүлж,
3. Монгол улсын их сургуулийн (МУИС), Шинжлэх ухааны сургуулийн Хүмүүнлэгийн ухааны салбарын Их Британи Америк судлалын тэнхимийн семинар 1 удаа,
4. МУИС-ийн Хэрэглээний шинжлэх ухаан, инженерчлэлийн сургуулийн их семинар, мөн сургуулийн Мэдээлэл, компьютерийн ухааны тэнхимийн семинарт 2 удаа,
5. Монгол улсын Шинжлэх ухаан технологийн их сургуулийн Мэдээлэл, холбооны технологийн сургуулийн эрдмийн зөвлөлийн нээлттэй семинарт 1 удаа,

6. Монгол улсын Шинжлэх ухааны академийн Физик технологийн хүрээлэнгийн их семинарт 1 удаа,
7. Италийн Трэнтогийн их сургуулийн Мэдээллийн инженерчлэл, Компьютерийн ухааны тэнхимийн эрдэм шинжилгээний семинарт 1 удаа тус тус илтгэж,
8. Трэнтогийн их сургуулийн судалгааны KnowDive группэд 4 [30]–[33] эрдэм шинжилгээний тайлан бичиж,
9. KnowDive группээс зохион байгуулдаг, урилгаар оролцдог *Knowledge in Diversity* эрдэм шинжилгээний хуралд 2 илтгэл хэлэлцүүлсэн.

Энэ диссертаци оршил, 3 бүлэг, дүгнэлт, ном зүй, талархал, 6 хавсралт, 25 хүснэгт, 43 зураг, нийт 130 хуудастай. Диссертацийн **нэгдүгээр бүлэгт** семантик технологийн тухай товч өгүүлж, түүгээр мэдлэгийн сан үүсгэж бодит ертөнцийг загварчлах, тэр дундаа олон хэл соёлыг хамарсан Ертөнцийн мэдлэгийн санг (ЕМС) үүсгэхэд тулгарч буй ялгамжийн асуудлыг хэлний болон ойлголтын түвшинд дэвшүүлсэн. Мөн олон хэлний онтологийг нутагшуулах аргуудыг дүгнэж шийдвэрлээгүй асуудлыг тайлбарлав. **Хоёрдугаар бүлэгт** онтологийг нутагшуулахад олны хүчийг ашиглах боломжийг судалж түүнийг хэрэгжүүлэхэд шаардлагатай алхмууд, олны хүчийг ижил төстэй ажил буюу орчуулгад ашигласан туршлага, үр дүн, орчуулгын чанарын үнэлгээ болон олон нийтийн оруулсан хувь нэмрийг үнэлэх аргын талаар өгүүлэв. Энэ ажлаар боловсруулсан онтологийг олны хүчээр нутагшуулах аргачлал, түүний хэрэгжүүлэлт, туршилтын үр дүнг төгсгөлийн **гуравдугаар бүлэгт** үзүүлэв.

БҮЛЭГ 1. ОНТОЛОГИ БА ЯЛГАМЖИЙН АСУУДАЛ

Энэ бүлэгт семантик технологийн хөгжлийн чиг хандлага ба ирээдүйн хэрэгцээ шаардлага, хүн төрөлхтний мэдлэг, үйл явдлыг машинд загварчлах зарим оролдлогууд - мэдлэгийн сан ба бодит ертөнцийн загварчлалын тухай товч ойлголтыг багтаасан. Мөн ЕМС-ийн бүтэц, бүрдэл хэсгүүдийг танилцуулах бөгөөд олон хэлний онтологи, онтологи нутагшуулах аргууд, тулгардаг бэрхшээлүүдийн тухай, ялангуяа ялгамжийн асуудлыг дэвшүүлэн тайлбарлана.

1.1 Мэдлэгийн сан ба бодит ертөнцийн загварчлал

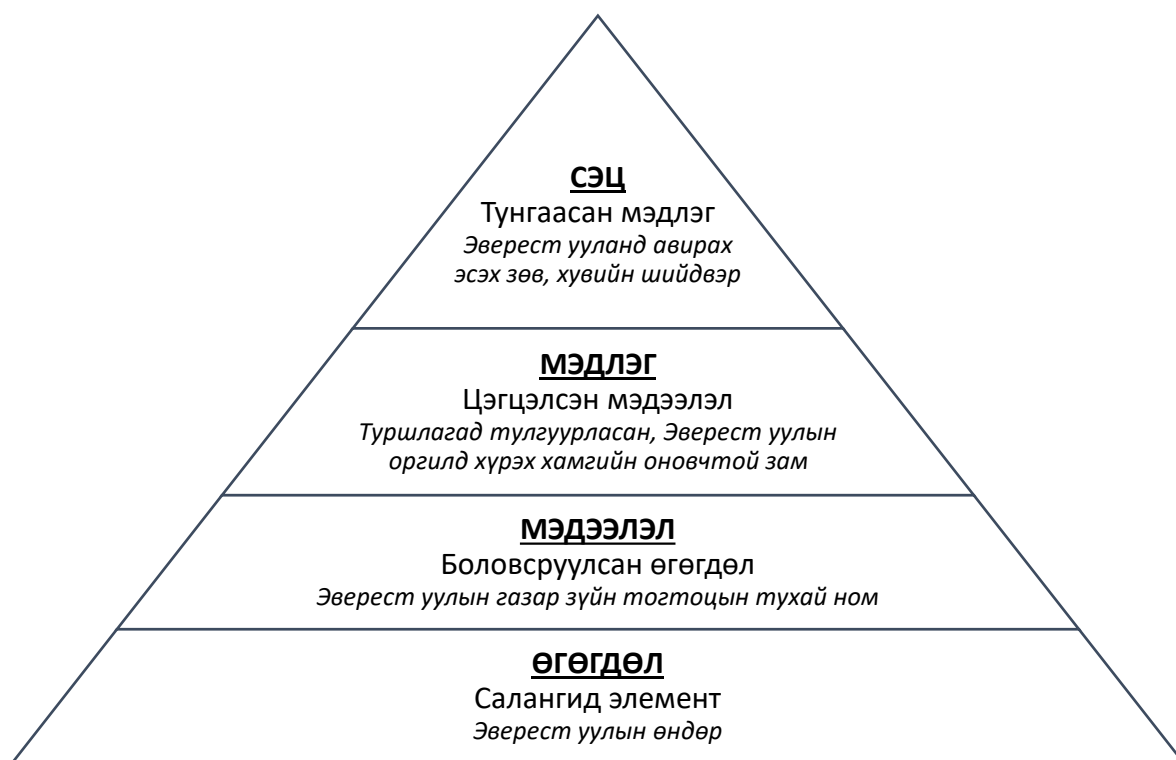
Семантик технологийн гол цөм бол мэдлэгийн сан бөгөөд онтологи нь түүний салшгүй нэг хэсэг байдаг. Энэ дэд бүлэгт семантик технологи, түүнд хэрэглэх мэдлэгийн сангийн тухай товч тайлбарлана.

1.1.1 Семантик технологи ба ирээдүй

Өнөөгийн ихэнх вэбүүд (WWW – World Wide Web) компьютерийн програмд зориулагдаагүй харин хүмүүс уншихад зориулсан материалууд, цахим баримт бичгүүдийг агуулсан байдаг. Хэдийгээр компьютер вэб хуудсыг задалж уншиж чадах боловч утгын түвшинд боловсруулалт хийхэд хүндрэлтэй. Жишээ нь, энэ бол Монгол улсын ерөнхийлөгч Ц.Элбэгдоржийн вэб хуудас байна, тэр холбоосоор түүний намтрыг үзэж болно гэх зэргээр машин ойлгож чадахгүй. Эрдэмтэн Тим-Бернерс Ли [1], семантик вэб бол машин ойлгоход амар утгат хэсгүүдийг агуулсан вэб хуудаснууд, нэг хуудаснаас нөгөөд чөлөөтэй шилжиж аливаа даалгаврыг хэрэглэгчид зориулан гүйцэтгэх програм хангамж ажиллах боломжтой орчныг үүсгэх дараа үеийн вэб [34] гэж тодорхойлсон байдаг. Энэ судалгаагаар ирээдүйд вэб хэрэглэгчийг төлөөлөх програмууд хоорондоо хамтран ажиллаж хүмүүст туслах тухай дурджээ. Ийм ухаалаг програм хангамж, роботууд вэб баримтаас өгөгдлийг ялган авч мэдээлэл боловсруулж ажиллах шаардлагатай. Гэтэл вэб дээрх өгөгдөл нь олон эх сурвалжийн, өөр хоорондоо ялгаатай тул өгөгдлийг нэгтгэх, нэгдсэн стандартыг маш сайн хэрэгжүүлэх асуудал 2009 оноос яригдаж эхэлсэн семантик вэбийн хөгжилд тушаа болсон хэвээр байна [2], [34].

Утгын түвшинд ажиллаж чадах програм хангамжийн системийн үндсэн технологи нь семантик технологи юм. Үүний цаана өгөгдлөөс мэдлэг гаргаж авах процессыг гүйцэтгэдэг (Зураг 1.1). Энд **өгөгдөл** гэдэг нь аливаа үйл явдал, баримтыг илэрхийлсэн, тодорхой бүтэцтэй бичвэр эсвэл тоон утгын цуглуулгыг гэж тодорхойлж

болно. Тодорхой бүтэцтэй гэдэг тухайн салангид бичвэр эсвэл тоон утгыг илэрхийлэх мета-өгөгдлийг хэлнэ. **Мета-өгөгдөл** гэдэг нь мэдээлэл эсвэл өгөгдлийн бүрдлийн бүтэц, формат, чанар, агуулга, бэлтгэн нийлүүлэгч, эх үүсвэр, боловсруулсан арга зүй, байрлалыг тодорхойлсон эсвэл тайлбарласан өгөгдөл юм. Иймд өгөгдлийг боловсруулж гарган авсан **мэдээллийг** аливаа харилцаа эсвэл өгөгдөл, баримтын талаарх мэдлэгийг илэрхийлж байгаа бичвэр, тоо, график, дуу, дүрс, газрын зураг зэргийг гэж ойлгоно.



Зураг 1.1 Өгөгдлөөс мэдлэг үүсэх шатлал

Харин семантик технологи агуулгын хүрээнд **мэдлэг** гэдэг бол туршлага эсвэл боловсролд тулгуурлан олж авсан цэгцэлсэн мэдээллийг, баримт сэлтийг хэлнэ. Хамгийн дээд түвшинд байрлах тунгаасан мэдлэгийг **сэц**¹ гэж тодорхойлов.

Ийм мэдлэгийг машинаар боловсруулж дэлхийн глобал асуудлаас байгууллага, хувь хүний явцуу асуудлыг, ухаалаг програм хангамжийн ч асуудлыг шийдвэрлэх боломжтой. Жишээлбэл, дэлхийн дулаарал, цаг уурын өөрчлөлт, авто замын түгжрэл, ухаалаг хотын шийдэл зэрэг асуудлуудыг өгөгдөлд түшиглэн шийдвэрлэж болно. Аливааг олон өнцгөөс харж өгөгдөлд түшиглэн шийдвэр гаргах үйл явц [35], илүү ухаалаг үйлчилгээ өнөөгийн нийгмийн бүхий л салбарт хэрэгцээтэй байна. Тухайлбал,

¹ Ой билиг, мэргэ оноц – Ш.Чоймаа нар, *Монгол хэлний хэлзүйн толь бичиг*, Улаанбаатар, 2005

хүн ам хурдацтай өсөж буй нөхцөлд хотын тээвэр өнөөдрийн томоохон хотуудад тулгарч буй нэг бэрхшээлтэй асуудал болжээ. Иймд байгаа нөөцөө илүү үр ашигтай, үр дүнтэй зарцуулж байгаль дэлхийтэйгээ зөв зохицож амьдрахын тулд өгөгдлийг маш ихээр цуглуулж түүнээс компьютерийн тусламжтайгаар мэдлэг гаргаж ашиглах шаардлагатай байна. Энэ хөгжилд хөтлөх гол технологиуд бол зүйлсийн интернет (Internet of Things), семантик технологи, хиймэл оюуны салбар юм.

Эдгээр технологиуд дундаас семантик технологийг анхлан практикт хэрэгжүүлж байгаа Фэйсбүүк², Гүүгл³ компаниуд өөрсдийн хайлтын бүтээгдэхүүний шинэ боломжуудыг хэрэглэгчдэд хүргээд эхэлжээ. Эдгээр хайлтын систем нь семантик технологийг ашиглан бодит нэгж-объектуудыг энгийн өгүүлбэрээс ялган тодорхойлж тэдгээрийн хоорондын холбоо хамаарлаар нь, шинж чанараар нь хайлт хийх боломжийг агуулсан байдаг. Жишээ нь, Гүүглийн хайлтын систем *Алберт Эйнштэйн хаана төрсөн бэ?* (*Where was Albert Einstein born?*) *Эйфл цамхгийн өндөр хэд вэ?* (*What is height of the Eiffel Tower?*), *Надад Трэнтогийн Пиацца Дуомо дээр авахуулсан миний зургуудаас үзүүлнэ үү?* (*Show me my photos at the Piazza Duomo in Trento*) зэрэг асуултуудад төвөггүй хариулна. Эхний жишээн дээр Эйнштэйн гэдэг хүнийг олоод түүний төрсөн газар болох Германы Улм хотыг хайлтын үр дүнд гаргаж ирэх бөгөөд энэ хүн тэр газар төрсөн гэдэг холбоосыг ашиглаж байгаа юм. Харин удаах асуулгаар Эйфлийн цамхгийн шинж чанарыг харуулах жишээтэй. Ийм ухаалаг үйлчилгээг үзүүлэхэд эх хэл боловсруулалтын аргаар өгүүлбэрийг задалж түүнээс бодит нэгж-объектыг хэдэн арван сая нэгж-объектууд, тэдгээрийн хоорондын холбоо хамаарлыг агуулсан мэдлэгийн сангаас хайж боловсруулснаар үр дүнг гаргадаг байна [36]–[38].

1.1.2 Мэдлэгийн сан

Мэдлэгийн цахим сан гэдэг аливаа зүйлсийн төрөл – класс (*гол, хот* г.м.), шинж чанар, тэдний хоорондын холбоо хамаарлыг (холбоос) Т хайрцагт (TBox), энэ хайрцагт байгаа зүйлийн бодит тохиолдлууд болох нэгж-объектууд (*Туул гол, Улаанбатаар* г.м.), тэдгээрийн шинж, хоорондын холбоо хамаарлыг А хайрцагт (ABox) тус тус хуваан хадгалдаг сан юм.

Мэдлэгийн сангийн ерөнхий бүтэц нь граф бөгөөд Зураг 1.2-т А, Т хайрцгийн жишээг үзүүлэв. Энэ санд мэдлэгийг дүрслэхдээ өгүүлэгдэхүүн (s) тусагдахуун (o)

² <http://search.fb.com/>

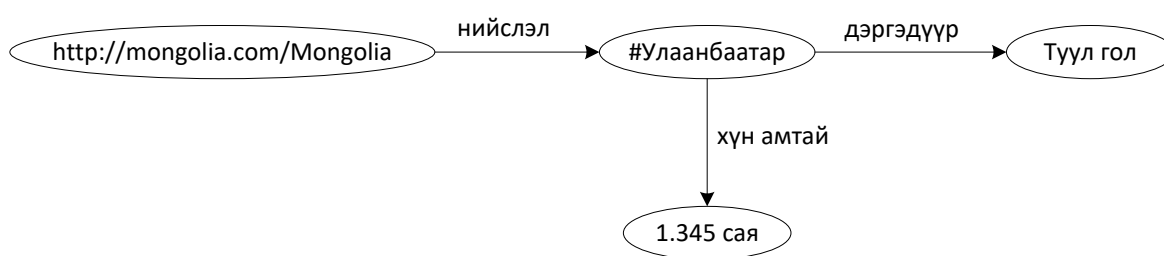
³ <https://www.google.com/intl/bn/insidesearch/features/>

өгүүлэхүүн (P) (subject-object-Predicate) гэсэн гурван элементийн нийлэмжээр илэрхийлдэг бөгөөд үүнийг гурвал гэнэ [3].

ТBox <u>ус бол газарзүйн байршил</u> <u>суурин газар бол газарзүйн байршил</u> <u>гол бол ус</u> <u>хот бол суурин газар</u> <u>нийслэл бол хот</u>	АBox хот (Улаанбаатар) гол (Түүл) улс (Монгол) нийслэл (Улаанбаатар, Монгол) дэргэдэх (Түүл, Улаанбаатар)
---	---

Зураг 1.2 Мэдлэгийн сангийн жишээ бүдүүвч

Жишээ нь, А хайрцагт мэдлэгийг *Улаанбаатар бол Монгол улсын нийслэл, Түүл гол Улаанбаатарын дэргэдүүр урсдаг, Улаанбаатар 1.3 сая хүн амтай* гэх мэт гурвалаар хадгалж болно. Үүнийг Зураг 1.3-д графикаар дүрсэлж үзүүлэв.



Зураг 1.3 Гурвалын график дүрслэл

Ийм гурвалыг W3C (World Wide Web Consortium) консорциумаас баталсан RDF⁴ (Resource Description Framework) стандарт загвараар хадгалдаг. Энэ стандарт нь XML хэл дээр түшиглэсэн байдаг. Зураг 1.3-д үзүүлсэн эхний жишээг RDF форматаар Зураг 1.4-т үзүүлсэн шиг илэрхийлж болно.

```

<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:mydomain="http://www.mydomain.org/my-rdf-ns">
  <rdf:Description rdf:about="http://mongolia.com/Mongolia">

```

⁴ <https://www.w3.org/RDF/>

```

<mydomain:нийслэл rdf:resource="#Улаанбаатар"/>

</rdf:Description>

</rdf:RDF>

```

Зураг 1.4 RDF жишээ

Дээрх зурагт (x, y, P) гурвалыг $P(x, y)$ логик томъёогоор илэрхийлж болох бөгөөд энд P бол x болон y нэгж-объектыг холбож буй хоёртын холбоос юм. Эдгээр бүх объектууд URL (Uniform Resource Locator)-ээр тодорхойлогдоно.

```

(#Улаанбаатар,
http://mongolia.com/Mongolia,
http://www.mydomain.org/нийслэл)

```

Өнөөдөр ихэнх мэдлэгийн сангууд энэ стандартаар хоорондоо өгөгдөл солилцох замаар холбогдсон байдаг. Хамгийн өргөн агуулгатай түгээмэл мэдлэгийн сангууд бол DBPedia [4], YAGO [6], WikiData [5], OpenCyC [39], Freebase [36]–[38] юм. Эдгээр сангуудыг М.Фарбер [40] нар өгөгдлийн чанарын 34 үзүүлэлт тодорхойлж чанар, агуулгын хамрах хүрээ, үнэн зөв байдлаар нь нарийвчлан харьцуулжээ. Энэ судалгаагаар мэдлэгийн сангийн агуулгын хамрах хүрээ, багтаамж (Хүснэгт 1.1) харилцан адилгүй, бүгд өөр өөрийн гэсэн онцлог шинжүүдтэй тул тохирох мэдлэгийн санг авч ашиглах нь зүйтэй гэж үзсэн байна.

Хүснэгт 1.1 Түгээмэл мэдлэгийн сангуудын агуулгын тоон харьцуулалт [40]

Элемент/Мэдлэгийн сан	DBPedia	Freebase	OpenCyc	Wikidata	YAGO
Гурвал	411M	3.12B	2.41M	748M	1.00B
Класс	736	53.0K	116K	302K	569K
Холбоос	58.7K	70.9K	18.0K	1.87K	106
Ялгаатай өгүүлэхүүн	60.2K	784K	165	4.83K	88.7K
Нэгж-объект	4.29M	49.9M	41.0K	18.6M	5.13M

В – тэр бум, М – сая, К – мянга

DBPedia бол ВикиПедиагийн⁵ бүтэцлэгдсэн мэдээлэл – товч хүснэгт (infobox), өгүүллийн ангилал, газарзүйн координат болон бусад гадаад линк зэргээс автоматаар гаргаж авсан мэдлэгийн сан юм. Энэ сан мэдлэгийн бусад сангуудтай хэдэн зуун сая

⁵ <https://www.wikipedia.org/>

холбоосоор холбогдсон байдаг. Тухайлбал, газарзүйн нэрийн мэдээллийн сан GeoNames⁶, АНУ-ын тагнуулын төв газрын CIA World Factbook⁷, хөгжмийн мэдээллийн MusicBrainz⁸ зэрэг олон төрлийн сангуудтай холбоотой өгөгдлийг хадгалдаг. Энэ сан өгөгдлийг илэрхийлэх нэрс, өгүүллийн нэр, товч тайлбар, өгүүллийн гадаад хэл дээрх линк зэргийг 13 хэл дээр хадгалдаг.

YAGO ч бас ВикиПедиагийн бүтэцлэгдсэн мэдээллээс гаргаж авсан мэдлэгийн сан юм. Гэвч АНУ-ын Принстоны их сургуулиас хөгжүүлсэн үгийн утгазүйн сүлжээ (lexical semantic network) ВёрдНэтийг⁹ ВикиПедиагийн өгөгдөлтэй нэгтгэж гаргасан анхны оролдлого бөгөөд ВёрдНэтийг ашиглан мэдлэгийн сангийн классыг тодорхойлсон байна. YAGO нэгж-объектын нэрс, товч тайлбар зэргийг 326 хэл дээр хадгалдаг.

WikiData бол ВикиПедиаг хөгжүүлдэг ВикиМедиа холбооноос санаачилсан, харьцангуй сүүлд, 2012 оны сүүлээр үүссэн мэдлэгийн сан юм. Энэ сан ВикиПедиагийн мэдээллийг хадгалдаг бөгөөд нэмж өгөгдлийн үнэн зөв байдлыг нотлох эх сурвалжуудыг давхар хадгалдаг. Нэгж-объектын нэрс, товч тайлбар зэргийг 400 гаруй хэл дээр хадгалдаг. Бусад мэдлэгийн сангаас ялгарах бас нэг онцлог шинж нь ВикиПедиа шиг хэрэглэгчид өгөгдлийг чөлөөтэй нэмж засварлаж болдог. Үнэндээ Гүүгл компанийн хөгжүүлсэн байсан Freebase мэдлэгийн сан олны хүчийг ашиглан өгөгдлийг баяжуулж сайжруулдаг сан байсан юм. Гэвч Гүүгл компани 2015 оны 6 дугаар сард энэ мэдлэгийн сангаа хөгжүүлэлтээ албан ёсоор зогсоож WikiData сан руу нэгтгэсэн билээ [36].

1.2 Ертөнцийн мэдлэгийн сан

UKC (Universal Knowledge Core) [41] бол Италийн Трэнтогийн их сургууль дээр хөгжүүлж буй мэдлэгийн сан юм. Энэ нь бодит ертөнц дээр нэгж-объектууд, тэдгээрийг илэрхийлэх хэдэн зуун ойлголтыг агуулсан бөгөөд *айн цөм (domain core)*, *ойлголтын цөм (concept core)*, *хэлний цөм (natural language core)* гэсэн гурван үндсэн бүрдлээс тогтоно (Зураг 1.5).

⁶ <http://www.geonames.org/>

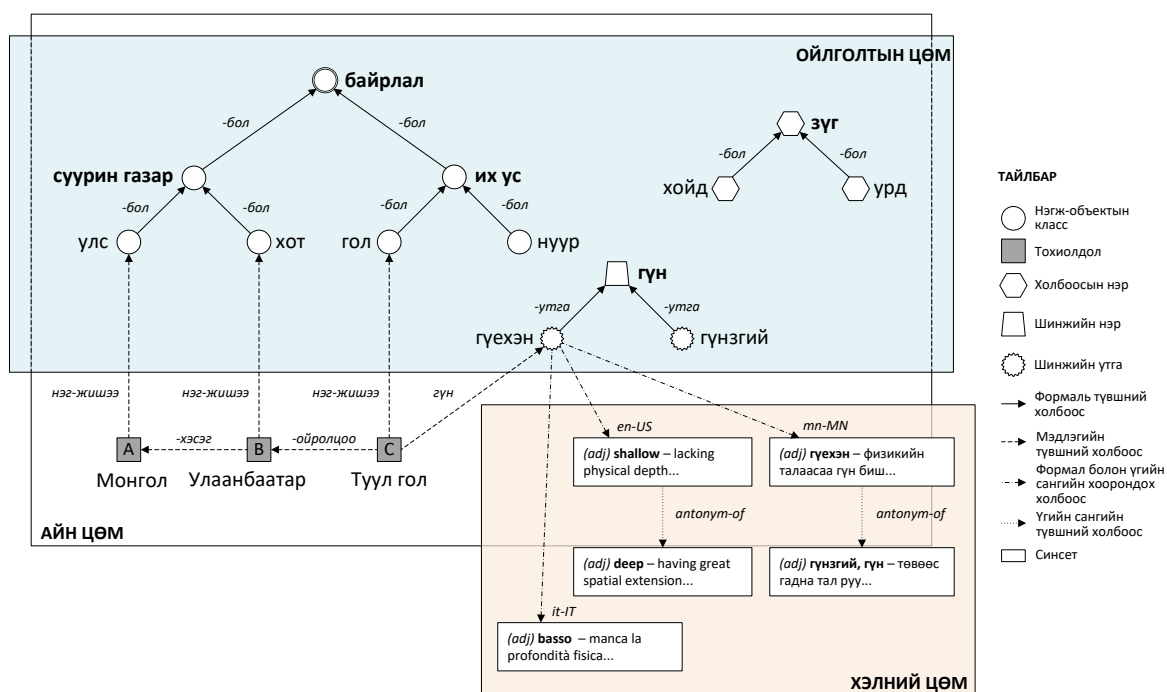
⁷ <https://www.cia.gov/library/publications/the-world-factbook/>

⁸ <http://musicbrainz.org/>

⁹ <http://wordnet.princeton.edu/>

1.2.1 Айн цөм

Айн цөм нь олон төрлийн айгаас тогтох бөгөөд ай бүр бидний сонирхож буй эсвэл бид хоорондоо харилцахдаа ашигладаг ямар нэг мэдлэгийн салбар эсвэл судлах талбарыг бүхэлд нь илэрхийлдэг [42]. Өөрөөр хэлбэл, ай бол судлагдахууны нийтлэг сэдэв (жишээ нь, математик, физик), мэргэжлийн салбарын хэрэглээ (жишээ нь, инженер, уул уурхай), тэдгээр зүйлсийн нийлэмж (жишээ нь, физикийн шинжлэх ухаан, нийгмийн шинжлэх ухаан) эсвэл өдөр тутмын амьдралын сэдэв (жишээ нь, спорт, хөгжим) байж болно. Ай нь фасетуудаас тогтох ба фасетийг аливаа мэдлэгийг олон талаас нь тодорхойлох төрөл ойлголтуудын шатлал (hierarchy) гэж тодорхойлж болно [13].



Зураг 1.5 Ертөнцийн мэдлэгийн сангийн бүтэц

DERA [12], эхний үсэг D нь Domain буюу ай, гэх аргачлалын дагуу фасетийг 3 ангилдаг: E – Нэгж-объектын класс (Entity class), R – Холбоос (Relation) болон A – Шинж (Attribute). Жишээ нь, орон зайн айн хувьд, *улс* болон *хот* бол нэгж-объектын класс, тэдгээр нэгж-объектуудыг хоорондох хамаарлыг тодорхойлох *ойролцоо*, *урд*, *цаана* гэдэг нь холбоосууд юм. Шинж бол нэгж-объектын шинж чанарыг илэрхийлдэг.

1.2.2 Ойлголтын цөм

Ойлголтын цөм нь ойлголтууд, тэдгээрийн хоорондын утгазүйн холбоосыг агуулдаг. Энэ цөмийн бүтэц нь циклгүй чиглэлт граф хэлбэртэй байдаг. Өөрөөр

хэлбэл, графын нэг орой (vertex) нь нэг ойлголт байх бөгөөд тал (edge) нь доороос дээш эсвэл доороос дээш эрэмбэтэйгээр хоёр оройг холбодог. Мөн энэ төрлийн граф төгсгөлөг шинж чанартай байна. Жишээ нь, Зураг 1.5-д location → populated place → city гэсэн чиглэлтэйгээр графыг байгуулж байна. Энд тал нь ойлголт хоорондын семантик холбоосыг илэрхийлнэ. Ийм графаар фасетийн бүтцийг, түүнд хэрэглэх үгсийг (ойлголтыг илэрхийлэх хэллэг) үүсгэдэг. DERA аргачлалын нэгж-объектын класс, холбоос, шинжүүд бүгд ойлголтоор илэрхийлэгдэнэ. **Ойлголт** гэдэг бол хэлний өгөгдсөн нэг үгтэй ойролцоо утгатай үгсийн олонлогийг илэрхийлэх хэлнээс үл хамаарсан илэрхийлэл юм. Жишээ нь, англи хэл дээр *city* гэдэг үгтэй ойролцоо утгатай *metropolis*, *urban center* үгсээр *хот* гэдэг ойлголтыг илэрхийлж болно. Мөн энэ ойлголтыг италиар *città* (chit'a), Монголоор *хот* гэдэг үгээр илэрхийлж болно.

Утгазүйн холбоос нь хоёр ойлголтын холбох холбоос юм. Жишээ нь, *is-a* (эсвэл *hyponym-of*), *part-of* (эсвэл *part-meronym-of*), *value-of* зэрэг холбоосууд байдаг. *is-a* холбоосны нэг жишээг дурдвал, хот *бол* (*is-a*) суурин газар. ЕМС-ийн холбоосуудыг Хавсралт А. Ертөнцийн мэдлэгийн сангийн холбоосноос дэлгэрэнгүй үзнэ үү.

1.2.3 Хэлний цөм

Хэлний цөм нь олон хэлнээс тогтох бөгөөд нэг хэл хэлзүйн объектууд, тэдгээрийн хоорондын холбоосуудаар илэрхийлэгдэнэ. Нэг объект үг (word), түүний утгалбар (sense), синсет (synset) эсвэл үгийн гажилттай хэлбэр (exceptional form) гэх зэрэг байж болно. **Үг** гэдэг нь хэлний үглэврээр илэрхийлэгдэх энгийн үгийн сангийн нэгж юм. Энэ нь тогтвортой нийлэмж - өвөрмөц хэллэг, хэвшмэл хэллэг зэрэг байж болдог. Энд **үглэвр** (lemma) гэдэг нь зөвхөн үгийн гадаад хэлбэр буюу үгийн бичигдэх, дуудагдах байдлыг хэлнэ [43]. Нөгөө талаас үг гэдэг нь ойлголтын цөм дэх ойлголтуудыг тухайн хэл дээр буулгасан орчуулга гэж ойлгож болно.

Үгийн утгалбарыг нэр, үйл, тэмдэг нэр, дайвар зэрэг 4 үгийн аймгаар (part-of-speech) зохион байгуулж хадгалдаг. Нэг үг нэгээс олон үгийн аймагт багтаж болох ижил үгийн аймагт байх ойролцоо утгатай үгийн утгалбарууд нэг синсет болж бүлэглэгдэнэ. Үгийн **утгалбар** гэдэг нь зөвхөн агуулга, утга болно [43]. Нэг үг нэг буюу түүнээс олон утгалбартай байж болох бөгөөд утгалбар бүр үгийн аймгийн тэмдэглэгээтэй байж болно. Нэг утгалбар нэг синсетэд хамаардаг. Тухайн үгийн бүх утгалбар хамгийн өргөн хэрэглэгддэгээрээ эрэмбэлэгдсэн байна. **Синсет** бол ойролцоо утгаар хэрэглэж болдог үгсийн олонлог. Үнэндээ синсетэд байгаа үгс утгын хувьд ойролцоо гэсэн холбоостой юм. Синсет бүр түүний утгыг тайлбарласан тайлбар,

утгыг тодотгосон жишээ өгүүлбэртэй байж болно. Хэлний цөм дэх холбоосууд нь үгзүйн (lexical) болон үг-утгазүйн (semantic lexical) зэрэг 2 төрөлд багтдаг. Эдгээр холбоосууд зөвхөн тухайн хэлний объектууд хооронд л байдаг.

Үгзүйн холбоос өөр синсетүүдийн үгсийн хооронд үүсэх холбоос юм. *Эсрэг үг (antonym)*, *үүсмэл-холбоотой-хэлбэр (derivationally-related-form)*, *бас харах (also-see)* зэрэг холбоосууд бол энэ төрлийн холбоосны жишээ болно. Эсрэг үг холбоосны доорх жишээнд давхар дундуур зураас өмнө синсетийн үгсийн таслалаар зааглаж, араас нь налуу форматаар синсетийн тайлбарыг, хаалтад жишээ өгүүлбэрийг үзүүлэв.

а) нам дор газар, хотгор газар -- *эх газартай холбоотой нам дор газар*

б) өндөрлөг газар, эх газрын өндөрлөг -- *эх газарт үүссэн өндөрлөг тогтоц*

Синсет (а)-д байгаа хотгор газар гэдэг үг бас синсет (б)-д байгаа өндөрлөг газар гэдэг үгтэй эсрэг үг холбоосоор холбогдоно. Энд *хотгор газар* гэдэг үг нь *эх газрын өндөрлөг* гэдэг үгтэй эсрэг үг холбоосоор холбогдохгүй. Учир нь үгзүйн холбоос зөвхөн тэдгээр үгсийн утгалбар хооронд үүсдэг.

Үг-утгазүйн холбоос бол хоёр синсетийн хооронд байдаг холбоос юм. Энэ холбоосны жишээ гэвэл, *төстэй (similar-to)*, *нэг-хэлбэр (troponymy)*, *үйлийн бүлэг (verb-group)* зэрэг юм. Үг-утгазүйн төстэй холбоосны жишээг дор үзүүлэв.

в) зэргэлдээ -- *хажууханд буюу маш ойр боловч салангид ориших (уулын зэргэлдээ газар; Нью-Йорк ба зэргэлдээ хотууд)*

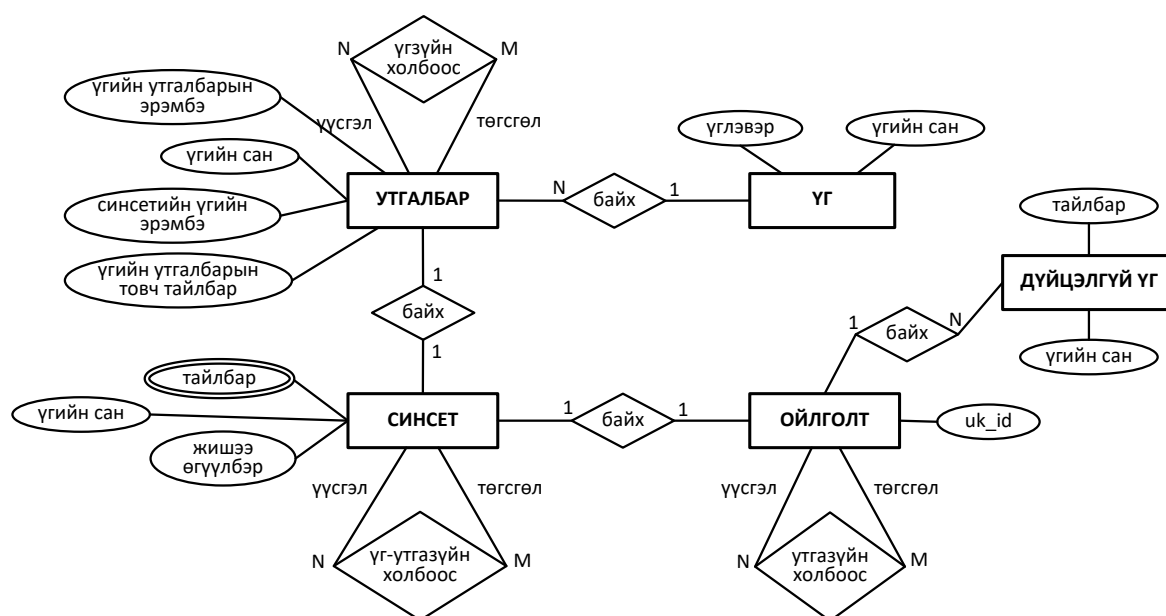
г) ойр, ойрхон -- *цаг хугацаа, орон зай, нөхцөл байдал зэргээрээ хол биш байдал (хажуугийн хөршүүд; ойрын ирээдүйд; тэр амжилтад хүрэхэд тун ойрхон байна; маш их ойр зүйл; бөмбөгөөр ойрхон цохих; тэр эмэгтэй галзуурахад ойрхон байсан; ойрын дуудлага байсан)*

Энэ тохиолдолд синсет (в) ганцхан үгтэй бөгөөд 3 үг агуулсан (г) синсеттэй *төстэй* холбоосоор холбогдож байна. Энэ нь синсет (в)-ийн ямар ч үг синсет (г)-ийн аль ч үгтэй ижил холбоосоор холбогдож болно гэсэн үг юм.

Хэлний цөм нь шаталсан зохион байгуулалттай синсетийн сангуудыг, тухайлбал, Принстоны их сургуулиас хөгжүүлсэн ВердНэт, Италийн FBK институтээс хөгжүүлсэн MultiWordNet¹⁰-ийн итали хэсгийг бүрэн (синсет нэг бүрийг

¹⁰ <http://multiwordnet.fbk.eu>

харгалзуулан) нэгтгэсэн. Хэлний цөмийн өгөгдлийн ойлголтын загварыг Зураг 1.6-д Нэгж-объект холбоосны диаграмаар харуулав. Энд нэг ҮГ олон УТГАЛБАР-тай байж болох ба үглэвэр шинжтэй байна. Эсрэгээрээ нэг УТГАЛБАР нэг ҮГ-тэй холбоотой. Утгалбарууд хоорондоо үгзүйн холбоосоор гэдрэг холбоос үүсгэх ба үүсгэл болон төгсгөлөөр холбоосны чиглэлийг зааж өгөх юм. Нэг утгалбар хэдэн ч утгалбартай үгзүйн холбоосоор холбогдож болно. Мөн утгалбар холбоостой үгийнхээ хэд дүгээр утга утга болохыг заах үгийн утгалбарын эрэмбэ, синсет дэх үгийн эрэмбэ, үгийн утгалбарын товч тайлбар зэрэг шинжүүдтэй болно. Тухайн үг ай дотор их/түгээмэл хэрэглэгддэг үгийн утгалбарын эрэмбэ өндөр байх бол бага хэрэглээтэй нь бага эрэмбэтэй байна. Синсет дэх үгийн эрэмбэ нь тухайн ижил утгатай үгсээс хамгийн их хэрэглээтэй нь өндөр, бага хэрэглээтэй нь бага эрэмбэтэй байна.



Зураг 1.6 Хэлний цөмийн нэгж-объект холбоосны (ERD) диаграм

Нэг УТГАЛБАР нэг СИНСЕТ-ээр илэрхийлэгдэх бөгөөд эсрэгээрээ нэг синсет нэг л утгалбартай байж болно. Мөн синсет, ойлголт хоёрын хооронд 1:1 харьцаатай байна. Харин синсетүүд өөр хоорондоо үг-утгазүйн холбоосоор M:N харьцаатай холбогдоно. ОЙЛГОЛТ-ууд хоорондоо утгазүйн холбоосоор холбогдох бөгөөд өөр хоорондоо дахин давтагдахгүй uk_id (universal knowledge identifier) шинжийг агуулна. Энэ нь хэлнээс үл хамаарсан байдаг. Ойлголтоос бусад нэгж-объектын төрлүүд бүгд үгийн сангийн нэгж гэсэн шинжтэй. Энэ нь тухайн нэгж-объект аль хэлний үгийн сангийн нэгжид багтахыг заах үүрэгтэй юм. Үлдсэн ДҮЙЦЭЛГҮЙ ҮГ нэгж-объектын төрлийн тухай тайлбар дараагийн дэд бүлгээс тодорхой болно.

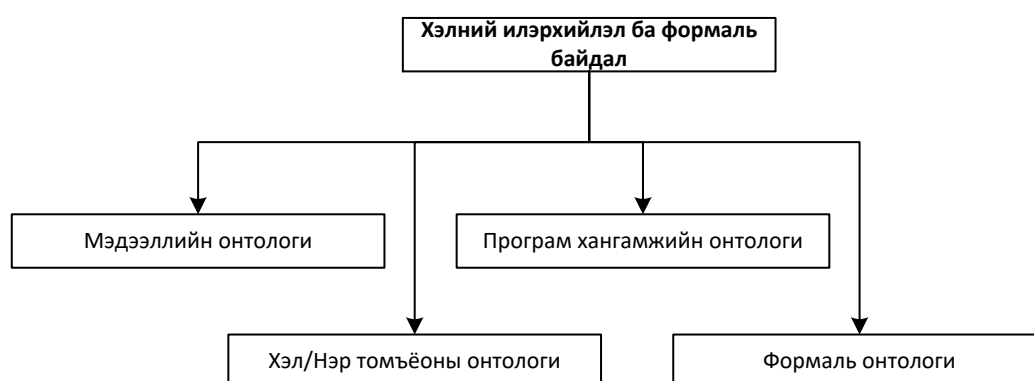
ЕМС 3 үндсэн бүрдлээр орчлон ертөнц дээрх бүх ойлголтыг аливаа хэл соёлоос үл хамаарах байдлаар нэгтгэн хадгалж түүнд хүний хэлээр хандах боломжийг хангаж өгсөн. Дэлгэрүүлбэл, ойлголт нь хэл соёлоос үл хамаарах орчлон дээр аль нэг соёлд тусгайлан хэрэглэгддэг эсвэл олон соёл хооронд нийтлэг хэрэглэгддэг бодит болон бодит бус зүйлийг илэрхийлэх мэдлэгийн бүрдүүлбэр нэгж бөгөөд түүнд аль ч хэл соёлоос хандах боломжийг олгож буй хэрэг юм. Харин ийм мэдлэгийн санг үүсгэх, зохион байгуулахад түүний цөм ойлголтуудыг агуулсан гол хэсэг онтологи буюу ойлголтын цөмийн олон хэл соёлын ялгаатай, нэг нэгнээсээ ялгарах өвөрмөц байдлыг тусгах асуудлыг авч үзэх ёстой.

1.3 Олон хэлний онтологи нутагшуулалт

Онтологи нутагшуулалтын тухай өгүүлэхээс өмнө онтологи гэж юу болохыг тайлбарлая. Дараа нь онтологи нутагшуулалт болон олон хэлний онтологийн тухай өгүүлнэ.

1.3.1 Онтологи

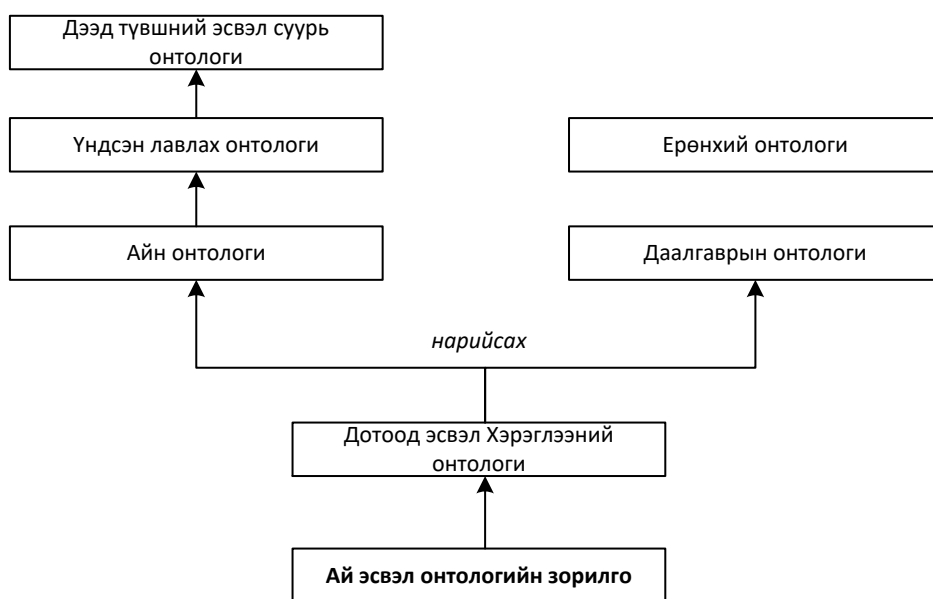
Олон эх сурвалжаас нэгтгэн товч ойлгомжтой хэлбэрээр тодорхойлбол, онтологи гэдэг нь хоорондоо утгазүйн холбоос бүхий ойлголтуудын олонлог юм [16], [44]. Дэлгэрүүлбэл, онтологи нь ойлголтууд, тэдгээрийн хоорондох утгазүйн холбоос, ойлголтыг илэрхийлэх үгийн сан, ойлголтын тайлбар, онтологийг бүхэлд нь тайлбарласан баримт бичгээс тогтдог. Онтологийг судлаачид цар хүрэнээс хамааруулж олон янзаар ангилах нь бий. К.Рөүсси нар [11] хэл болон формаль талаас (Зураг 1.7), онтологийн зорилго эсвэл айн (тодорхой хүрээг хамарсан сэдэв) талаас нь тус тус нэгтгэн ангилсан байна.



Зураг 1.7 Хэл болон формаль талаас ангилсан онтологийн төрөл [11]

Мэдээллийн онтологи (Information Ontology) нь төслийн гүйцэтгэлийн явцад оролцогчдын санаа бодлыг цэгцлэх, түүнд хэрэглэх ойлголтуудын нарийн зааг ялгааг

гаргахад ашигладаг диаграм болон зураглалуудаас бүрддэг. Тухайлбал, Mind Map төрлийн програмаар мэдээллийг сэдэвчилж шаталсан хэлбэрээр харуулж байгаа нь энэ онтологийн нэг жишээ юм. *Хэл/Нэр томъёоны онтологи* (Linguistic/Terminological Ontology) үгсийн түүвэр (glossary), толь бичиг, үгийн сан, таксономи (taxonomy), тайлбар толь эсвэл үгийн сангийн өгөгдлийн сан зэрэг байдаг. Өргөн хэрэглэгддэг SKOS (Simple Knowledge Organization System), RDF хэлнүүдийг ашиглан эдгээр онтологийг үүсгэж болдог. *Програм хангамжийн онтологи* (Software Ontology) програм хангамжийн хөгжүүлэлтэд хэрэглэгддэг, өгөгдлийн нийцлийг хангах зорилгоор ихэвчлэн өгөгдлийн хадгалалт, өгөгдлийн ашиглалтад анхаарсан ойлголтын схем байдаг. UML (Unified Modeling Language) стандартын класс болон объект диаграмууд ийм онтологийг ихэвчлэн тодорхойлдог. *Формаль онтологи* нь ойлголтыг ямар нэг илэрхийлэх хэл дээрх тодорхой утгазүй, ойлголт болон тэдгээрийн холбоосыг тодорхойлох тухай дүрмүүдийг шаарддаг. Үүнд формаль логикийг ашигладаг бөгөөд мэдлэгийн сан логик тодорхойлолтоор аливаа зүйлсийн утгыг илэрхийлдэг формаль систем юм.

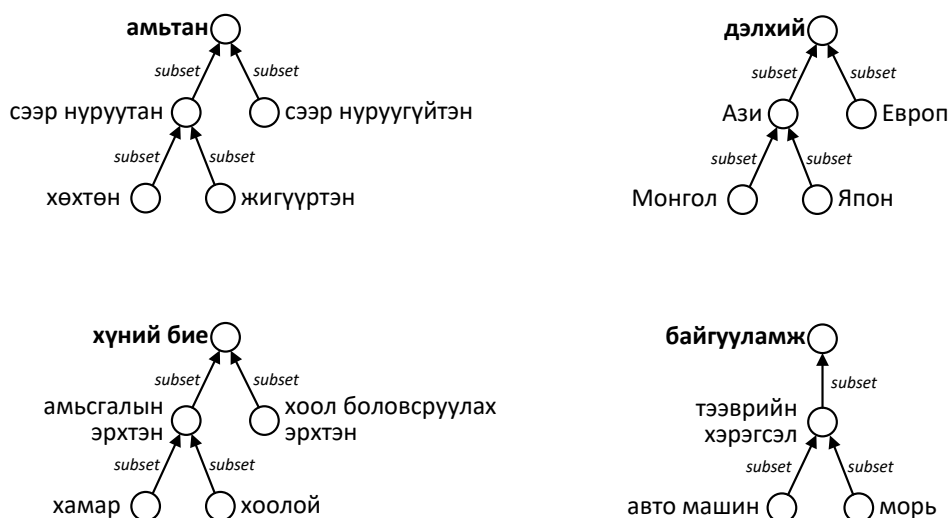


Зураг 1.8 Айн зорилгод түшиглэсэн онтологийн ангилал [11]

Формаль логикийг хэрэгжүүлдэг хэлнүүд гэвэл Description Logics (DL), Conceptual Graphs (CG), First Order Logic (FOL), OWL (Web Ontology Language) юм.

Онтологийн зорилго болон айн талаас ангилбал (Зураг 1.8), *Дотоод эсвэл хэрэглээний онтологи* (Local/Application Ontology) нь *айн онтологийн* (Domain Ontology) нарийвчилсан хувилбар бөгөөд айн тухайн нэг загварын дагуу зөвхөн нэг

хэрэглэгч буюу нэг хөгжүүлэгчийн зүгээс нэг харах өнцгөөс илэрхийлэх онтологи байдаг. Харин *айн онтологи* нь тодорхой хүрээллийн хүмүүс аливаа ойлголтыг хэрхэн төсөөлж, хэрхэн хүлээж авч буй өнцгөөс тодорхойлсон, нэг айд хамаарах ойлголтуудын онтологи байдаг. *Үндсэн лавлах онтологи* (Core Reference Ontology) [45] бол өөр өөр бүлгийн хүмүүс дундаа хэрэглэж болох стандарт онтологи тэдгээр хүмүүсийн харах ялгаатай өнцгүүдийг нэгтгэсэн байдаг. *Ерөнхий онтологи* (General Ontology) ямар нэг айтай холбоогүй, өргөн хүрээнд ерөнхий мэдлэгийг агуулсан байдаг. Харин *даалгаврын онтологи* (Task Ontology) нь ямар нэг даалгаврыг гүйцэтгэхэд шаардлагатай мэдлэгийг агуулдаг бол айн онтологи тэр даалгавартай холбоотой салбарын мэдлэгийг агуулдаг. *Дээд түвшний эсвэл суурь онтологи* (Top level or Foundational Ontology) нь олон айд хэрэглэж болох ерөнхий зориулалтын онтологи.



Зураг 1.9 Ангилал онтологийн жишээ [12]

Харин онтологийг хэлбэр, ойлголтуудын дундах холбоосны ерөнхий төрөл талаас Ангилал онтологи (Classification Ontologies) болон Дүрслэл онтологи (Descriptive Ontologies) гэж хоёр ангилж болно [12]–[14]. Ангилал онтологийн хувьд ихэвчлэн баримтыг (document) тодорхойлох, ангилах болон баримтаас хайлт хийхэд хэрэглэгддэг. Өөрөөр хэлбэл, баримтуудыг тодорхойлох нэр томъёо (term) нь BT (Broader Term)/ NT (Narrower Term) эсвэл superset/subset зэрэг шаталсан холбоосуудаар хоорондоо холбогдсон байдаг (Зураг 1.9). Жишээ нь, ангилал онтологийг ашигласан мэдлэгийг зохион байгуулах (Knowledge Organization) системээс морь гэдэг хөхтөн амьтны тухай хайхад морийг тээврийн хэрэгслийн болох

тухай баримтууд мөн олдоно. Энэ тохиолдолд хайлтын илэрцийн оновчтой үр дүнг ялгах нь хүндрэлтэй байдаг.

Мэдлэгийг дүрслэх (Knowledge Representation) системд дүрслэл онтологийг ашигладаг. Энэ нь бодит ертөнц дээр нэгж-объектуудыг тухай учир шалтгаан, зорилгод тулгуурлан үүссэн онтологи юм. Ийм онтологийн нэр томъёонууд нэгж-объектуудыг илэрхийлж шаталсан *энэ бол...* холбоосоор тэдгээрийг холбох маягаар онтологийн гол бүтцийг үүсгэдэг. Жишээ нь морь бол амьтан, бас тээврийн хэрэгсэл гэсэн холбоос энэ 3 ойлголтын дунд байх юм. Үүнийг бодит ертөнцийн утгазүйн холбоос гэнэ [13]. Энэ онтологийн жишээг Зураг 1.5-аас харж болно. Дүрслэл онтологийг ашигласан мэдлэгийг дүрслэх системд баримтууд нэгж-объектуудыг олноор агуулсан бүхэл зүйл юм. Иймд мэдлэгийн дүрслэл нь мэдлэгийн ангиллаас илүү нарийвчилсан түвшинд мэдээллийг зохион байгуулж хадгалдаг.

Онтологийн энэ олон ангиллаас үзэхэд онтологийг ямар талаас нь харахаас хамаараад өөр өөрөөр ангилан ялгаж болохоор байна. Ийм учраас нэг онтологи өөртөө олон шинжийг агуулсан байж болно. Жишээ нь 1.2 дугаар дэд бүлэгт тайлбарласан EMC-ийн ойлголтын цөм нь DL хэлээр бичигдсэн формаль бөгөөд дүрслэл онтологи, түүнд хандах хэлний цөм бол хэл/нэр томъёоны онтологи юм. Айн цөм бүрээр нь хуваан авч үзвэл айн онтологи, нэг айг тусгайлан нэг хэрэглээнд ашиглах бол хэрэглээний онтологи болно. Энэ мэдлэгийн санд бүх айд хэрэглэж болох ерөнхий ойлголтуудыг хадгалдаг тул дээд түвшний онтологийн шинжийг бас агуулна. Ийм сан өөртөө олон төрлийн онтологи агуулсан байх бөгөөд аливаа зорилгоор өргөн хүрээнд хэрэглэж болох зөв бүтэцтэй мэдлэгийн сан юм.

1.3.2 Онтологи нутагшуулалт

Онтологийг ихэвчлэн гар аргаар мэргэжлийн хүмүүсийн оролцоотойгоор, тэр дундаа сэтгэц хэл шинжлэлийн мэргэжилтнүүд хийдэг. Өөрөөр хэлбэл, хүмүүс тухайн хэл соёлд ямар ойлголтуудаар юуг төсөөлж харилцдагийг, хүний тархи хуримтлуулсан туршлагынхаа хүрээнд аливааг хэрхэн танин мэдэж хүлээн авч буйг судалдаг мэргэжилтнүүд байдаг. Нөгөө талаас салбар салбарын мэргэжилтнүүд тухайн салбарт хэрэглэж буй нэр томъёог илүү сайн мэддэг. Иймд энэ ажил мэргэжлийн олон уулзвар дээр хамтран ажиллах өндөр мэргэжилтнүүдийг шаарддаг билээ. Гэвч зарим улс орнууд хэдэн арван жил гар аргаар онтологийг хөгжүүлэн ашиглаж байна. Гэвч эдгээр ихэвчлэн англи хэл дээр үүссэн байдаг. Харин дэлхий

глобалчлагдаж буй өнөө үед олон хэл дээр онтологийг ашиглах хэрэгцээ шаардлага улам бүр нэмэгдэж байна.

1.3.2.1 Онтологи нутагшуулах аргачлал

Олон хэл дээр онтологийг үүсгэх оновчтой арга бол онтологийг нутагшуулах арга юм. Онтологийг нутагшуулах гэдэг нь мэдлэгийн зарим хэсгийг ямар нэг хэл соёлруу нийцүүлэн буулгах үйл явцыг хэлнэ [7], [16]. Онтологи нутагшуулалт бол зөвхөн өгөгдсөн онтологийн (эх) хүрээнд байгаа ойлголтуудыг гаралтын онтологид (зорилгын) буулгах бус гаралтын онтологийг шинэчлэн өөрчлөх гэсэн санааг давхар агуулж байдаг. Энэ арга өмнө нь үүссэн аль нэг эх онтологийн цөм бүтцийг ашиглан шинэ зорилгын онтологийг цоо шинээр үүсгэснээс харьцангуй хурдан, зардал багатайгаар үүсгэдэг боловч энэ нь цаг хугацаа, зардал, мэргэжлийн хүч шаардсан ажил хэвээр байгаа. Хэдий тийм боловч судлаачид энэ бэрхшээлийг даван туулахын тулд оновчтой арга замыг эрэлхийлсээр байна.

Эрдэмтэн М.Эспиноза нар онтологийг нутагшуулах хагас автомат аргыг [10] санаачлан хэрэгжүүлсэн байна. Энэ нь 1) машин орчуулгын системээр эх онтологийн нэр томъёог орчуулан улмаар 2) үгийн утга салгуурын WSD (Word Sense Disambiguation) тусламжтайгаар тухайн үгийн утгалбарыг тодорхойлдог. Үүний дараа 3) мэргэжилтнүүд хянаж баталгаажуулдаг. Түүний арга нь машин орчуулгын алхамдаа EuroWordNet [46], Google Translate¹¹, Wiktionary¹², Babelfish¹³ болон FreeTranslation¹⁴ зэрэг үгийн сангийн өгөгдлийн сан, машин орчуулгын үйлчилгээг ашигладаг. Мөн үгийн утга салгагч нь үг сангийн өгөгдлийн сан дотроо маш их ашигладаг. Дунджаар энэ арга 72%-ийн нарийвчлалтайгаар зөв нутагшуулдаг байна. Хэдийгээр энэ арга мэргэжилтний оролцоог буруулж байгаа боловч маш олон төрлийн нөөц, орчуулгын сайн үйлчилгээг ашигладаг. Бас М.Аркан нар статистик машин орчуулгын системийг өргөтгөн онтологийн нэр томъёоны утгалбарыг Англи-Герман зэрэгцээ материалын сангаас утгазүйн төсөөг боддог алгоритмуудаар тооцож онтологийг нутагшуулах автомат аргыг боловсруулсан байна [17]. Мөн С.Валтер нар M-ATOLL [19] гэх програм хангамжийн систем хөгжүүлж Wikipedia олон хэлний материалын сан, DBpedia мэдлэгийн санг ашиглан онтологийн үг сангийн мэдээллийг

¹¹ <https://translate.google.com>

¹² <https://www.wiktionary.org>

¹³ <https://www.babelfish.com>

¹⁴ <https://www.freetranslation.com>

хагас автоматаар гаргах аргыг боловсруулсан байна. Эдгээр аргыг дээрх нөөцөөр дутмаг, цахим нөөц багатай хэл соёлын хувьд ашиглах боломжгүй байна.

Судлаач Б.Лансер нар онтологийн үгийн санг (хэл/нэр томъёоны онтологи) англи хэлнээс япон хэл рүү олны хүчээр буюу олон хүмүүсийн оролцоотойгоор орчуулах замаар шинэ онтологи үүсгэх аргыг туршсан байна [22]. Мөн М.Бенжамин нөөц багатай олон хэлний толь бичгийг өгөгдлийн олборлолт, мэргэжилтний оролцоо, олны хүч болон тоглоомын аргыг нэгтгэн ашиглаж үүсгэх загварын санаагаа танилцуулсан байна [21]. Эдгээр аргууд нь цаг хугацаа, зардал шаардах гол асуудлыг шийдэж болмоор санаа боловч нарийн судлагдаагүй, дөнгөж эхний туршилтууд хийж тандсан төдийд байна.

1.3.2.2 Онтологи нутагшуулах давхарга

Онтологи нутагшуулах чиглэлээр тэргүүлэх судалгаа хийдэг эрдэмтэд П.Чимиано, Е.Монтиел-Понсода, П.Буйтлаар, М.Эспиноза нар онтологи нутагшуулалтыг үгийн сангийн (lexical layer) болон ойлгомжийн (conceptualization layer) гэсэн хоёр давхаргад хуваан авч үзсэн байна. Энэ нь ЕМС-ийн ойлголтын цөм, хэлний цөм хоёртой нэг талаараа дүйх ухагдахуун юм. *Үгийн сангийн давхарга* нь онтологи нутагшуулалтын явцад өөрчлөгдөх нь гарцаагүй. Өөрөөр хэлбэл, тухайн хэл соёлоос хамаарч өөр өөр үг хэрэглэгдэх болно. Энэ давхаргад тохиолдох хамгийн энгийн нутагшуулалт бол үг болон түүний тайлбарын 1:1 орчуулга юм.

Үгийг орчуулах нь онтологи нутагшуулах ажлын зайлшгүй хийх ёстой даалгавар. Харин *ойлгомжийн давхаргыг* шаардлагатай бол өөрчилдөг. Өөрөөр хэлбэл, эх онтологид байгаа ойлголт зорилтот хэл соёлд 1:1 харьцаагаар нийцэхгүй байж болно. Энэ тохиолдолд зорилгын онтологи нь ойлгомжийн түвшинд өөрчлөгдөн үүсэх юм.

Дээрх хоёр давхаргууд хоорондоо нягт холбоотой. Ойлгомжийн давхаргад өөрчлөлт ороход үгийн сангийн давхарга дагаад зайлшгүй өөрчлөгддөг. Харин үгийн сангийн давхаргад өөрчлөлт ороход ойлгомжийн давхаргад нөлөөлж болдог бас нөлөөлөхгүй ч байж болдог. Тухайлбал, эх онтологийн тухайн ойлголт зорилгын онтологид ижил байх ба энэ хоёр онтологи үгийн сангийн түвшинд өөр өөрөөр нэрлэдэг бол үгийн сангийн давхаргын өөрчлөлт ойлгомжийн давхаргыг өөрчлөхгүй байж болно гэсэн санаа юм. Иймд дараах асуудлуудыг тодорхойлжээ [16].

1. Ойлгомжийн давхаргын өөрчлөлт үгийн сангийн давхаргыг өөрчилдөг
2. Зорилгын онтологид утгын шилжилт үүсдэг

3. Үгийн сангийн давхарга ойлгомжийн давхаргыг өөрчилж болно

1.3.2.3 *Онтологи нутагшуулах хэмжүүр*

Онтологи нутагшуулах ажлын төрлийг тодорхойлох 3 үндсэн хэмжүүр байдаг [7], [16].

1. Стандартчилсан ай эсвэл соёлын нөлөөт ай

Зарим ай олон улсын түвшинд өгөгдлийг солилцох зорилгоор стандарт болсон байдаг. Ийм ай ихэвчлэн техникийн ай байдаг. Жишээ нь, инженер, санхүүгийн салбар стандарт (жишээ нь, ISO) үйл ажиллагаатай эсвэл тайлангийн стандарт (жишээ нь, санхүүгийн XBRL - eXtensible Business Reporting Language) мөрддөг. Ийм айн ойлголтууд өөр өөрийн хэл дээр тогтсон нэр томъёогоор илэрхийлэгддэг учир онтологийг нутагшуулахад ихэвчлэн 1:1 харьцаагаар буулгадаг. Харин бусад соёлын нөлөөнд автдаг айн хувьд, жишээ нь, татвар, хууль, улс төрийн систем зэргээрээ ялгарах төрийн удирдлагын ойлголтууд соёл хооронд нэлээн зөрүүтэй байдаг.

2. Функционал эсвэл баримтжуулах нутагшуулалт

Онтологийг юунд хэрэглэхээс хамаарч өөр өөр зорилгоор нутагшуулж болно. Нэг талаас эх онтологитой яг ижил зорилгоор зорилгын онтологийг нутагшуулж болно. Жишээ нь, төрийн удирдлагын айд цагаачлалын үйл ажиллагааг илэрхийлсэн онтологийг өөр хэл соёлд хэрэглэх бол тухайн хэл соёлд тохирсон ойлголтуудаар нутагшуулах ба эх онтологийн зорилго, агуулгыг алдагдуулж болохгүй. Иймд функционал нутагшуулалт хуучин онтологид түшиглэн шинэ онтологийг үүсгэх асуудал юм. Нөгөө талаас баримтжуулах нутагшуулалтын зорилго бол эх онтологийг өөр хэл соёлтой иргэд хэрэглэх юм. Жишээ нь, цагаачлалын айн онтологийг гадаадын иргэд хэрэглэх бол тэдний хэлээр ойлголтуудын утгыг илэрхийлж нутагшуулна. Энэ тохиолдолд шинэ ойлголтуудтай онтологи үүсэхгүй, зөвхөн эх ойлголтуудыг өөр хэлээр илэрхийлсэн онтологи үүснэ.

3. Нэгдмэл эсвэл салангид

Онтологийг нутагшуулах бас нэг чухал онцлог бол зорилгын онтологи эх онтологитой хэр зэрэг нэгдмэл, хоорондоо нягт холбоотой байдал юм. Хэрэв зорилгын онтологийг эх нийгэм (source community) болон зорилтот нийгэм (target community) хоёрын хооронд өгөгдөл солилцоход хэрэглэх бол ойлгомжийн давхаргад аль болох өөрчлөлт багатай байх хэрэгтэй. Ийм маягаар нэгдмэл

байдлыг хангадаг. Хэрэв зорилгын онтологи салангид байдлаар хэрэглэгдэх бол ойлгомжийн давхаргад шаардлагатай өөрчлөлтүүдийг оруулж зорилгын нийгмийн хэрэгцээ шаардлагыг хангах учиртай.

Эдгээр хэмжүүр онтологи нутагшуулалтын ямар давхаргад нөлөөлөхийг Хүснэгт 1.2-т үзүүлэв. Стандартчилсан айн онтологийг нутагшуулахад шинэ ойлголтуудтай онтологи үүсдэггүй, харин функционал зорилгоор онтологийг нутагшуулах үед шинэ онтологи үүсдэг тул энэ хоёрыг зэрэг хангасан онтологи байх боломжгүй гэж үзсэн байна.

Хүснэгт 1.2 Онтологи нутагшуулалтын төрөл [16]

Зорилго/Айн төрөл	Стандарт	Соёлын нөлөөт
Функционал	-	Үгийн сангийн давхарга, Ойлгомжийн давхарга
Баримтжуулах	Үгийн сангийн давхарга	Үгийн сангийн давхарга

1.3.3 Онтологи нутагшуулалтын асуудлууд

Онтологийг нутагшуулахад орчуулгын, менежментийн болон олон хэлний илэрхийллийн асуудлууд тулгардаг [7].

1. Орчуулгын асуудал

Нэгэнт аливаа хэл соёл бүхэн өөр өөр үгийн сантай учир онтологийн нэр томъёог орчуулах үед дараах асуудлууд тулгардаг.

i. Яг ижил нэр томъёоны олдоц

Энэ нь ихэвчлэн стандартчилсан онтологи нутагшуулалтын явцад тохиолддог бөгөөд шууд 1:1 үгээр эх онтологийн нэр томъёог буулгах юм. Жишээ нь, Монгол хэлний *спорт* гэдэг нэр томъёо англи хэлэнд *sport*, испани хэлэнд *deportes* гэж орчуулагдана.

ii. Агуулгаас хамаарсан хэд хэдэн нэр томъёоны сонголт

Аливаа хэлний нэг нэр томъёо өөр хэлний олон нэр томъёогоор орчуулагдаж болно. Энэ тохиолдолд онтологийн ерөнхий агуулга, айгаас хамаарч хамгийн тохиромжтой нэр томъёог сонгох шаардлагатай. Жишээ нь, англи хэлний *accommodation* гэдэг нэр томъёо испани хэл дээр *alojamiento* (Испани улсад), эсвэл *hospedaje* (Өмнөд Америкт) гэж орчуулагдаж болно. Орчуулга бүр

тухайн ойлголтын нарийн ялгааг илэрхийлэх бөгөөд тухайн онтологид хамгийн оновчтой эсвэл хамгийн ойр дүйх нэр томъёог сонгох хэрэгтэй.

iii. Ойлгомжгийн зөрүүтэй байдал

Нэг соёл нэг ойлголтыг тодорхой түвшинд илэрхийлэх, өөрөөр хэлбэл, тодорхой өнцгөөс харж тайлбарладаг бол тэр ойлголтыг өөр соёл арай өөр өнцгөөс ойлгодог бол ойлголтын ялгаатай байдал үүснэ. Энэ тохиолдолд тухайн ойлголтыг зорилгын онтологид дүйх нэр томъёогоор илэрхийлж чадахгүй тохиолдол үүснэ. Жишээ нь, франц соёлд далайд цутгадаг голыг *fleuves*, өөр голд цутгадаг голыг *rivières* гэдэг нэр томъёогоор тусад нь ялгаж илэрхийлдэг байхад англид энэ ялгааг гаргаж хэрэглэдэггүй.

2. Менежментийн асуудал

Онтологи нутагшуулахад явцад нэр томъёог нэг нэгээр орчуулагдахгүй тохиолдол олон байх тул үүнээс үүдэн зорилгын онтологийн нэр томъёог шинэчлэх, нийцүүлэн буулгах менежментийн асуудлыг авч үзэх ёстой. Гол бэрхшээл нь онтологийн нэр томъёо, тэдгээрийн орчуулгын өөрчлөлтийг зохион байгуулах юм. Энэ тохиолдолд дараах нөхцөл байдал үүсэж болно.

i. Онтологийн нэр томъёо шинээр нэмэгдэх

Энэ үед онтологийн нэр томъёо бүх хэл дээр орчуулагдаж бууна.

ii. Онтологийн нэр томъёо устаж алга болох

Энэ тохиолдолд бүх хэл дээрх орчуулгууд дагаад устах ёстой.

iii. Онтологийн нэр томъёо өөрөөр нэрлэгдэх

Ийм үед үүний орчуулга бүх нэр томъёонууд дахин хянагдана.

3. Олон хэлний илэрхийллийн асуудал

Онтологи нутагшуулалтын үр дүнд зорилгын онтологи нь өөр хэл дээр илэрхийлэгдэх нэр томъёонуудтай болно. Онтологийг олон хэлээр илэрхийлэхэд дараах асуудлуудыг тодорхойлж болно.

i. Олон хэлний мэдээллийг оруулах

Онтологи инженерчлэлийн нийгэмлэг, байгууллагууд олон хэлний мэдээллийг илэрхийлэх стандартуудыг ашигладаг. Жишээ нь, RDFS (Resource Description Framework Schema) стандартын *rdfs:label*, *rdfs:comment* шинжүүдийг ашигладаг.

ii. Нэг хэл соёлд шинэ ойлголт оруулах, тохирох ойлголттой дүйцүүлэн буулгах

Ойлголт бүр тухайн хэл соёлд өөрийн гэсэн утгыг илэрхийлдэг боловч бусад хэл соёлд үл ялиг зөрөөтэй байдлыг үл харгалзан дүйцүүлэн холбоно.

iii. Бусад олон хэлний мэдээллийг холбох

LIR (Linguistic Information Repository), LigInfo, LexOnto зэрэг олон хэлний мэдээллийн сангуудтай холбож онтологийн олон хэлний мэдээллийг баяжуулна.

ЕМС орчлон ертөнц дээрх ялгаатай болон нийтлэг бүх ойлголтуудыг агуулах зорилготой. Өөрөөр хэлбэл, олон улсын стандарт нэр томъёогоор илэрхийлэх ойлголтоос эхлүүлээд зөвхөн тухайн хэл соёлд хэрэглэх ойлголтуудыг хүртэл бүгдийг нь хамарсан мэдлэгийн сан юм. Ийм учраас дээрх онтологи нутагшуулалтын төрлүүдийг бүгдийг нь нэг дор багтаасан ерөнхий нэг аргачлал тодорхойлох зайлшгүй шаардлагатай юм. Бидний мэдэж байгаагаар энэ асуудлыг шийдсэн аргачлал байхгүй байна. Ялангуяа олон хэл соёл хоорондын онцлог, ялгааг нарийн тусгаж зохион байгуулах асуудал шийдэгдээгүй байна.

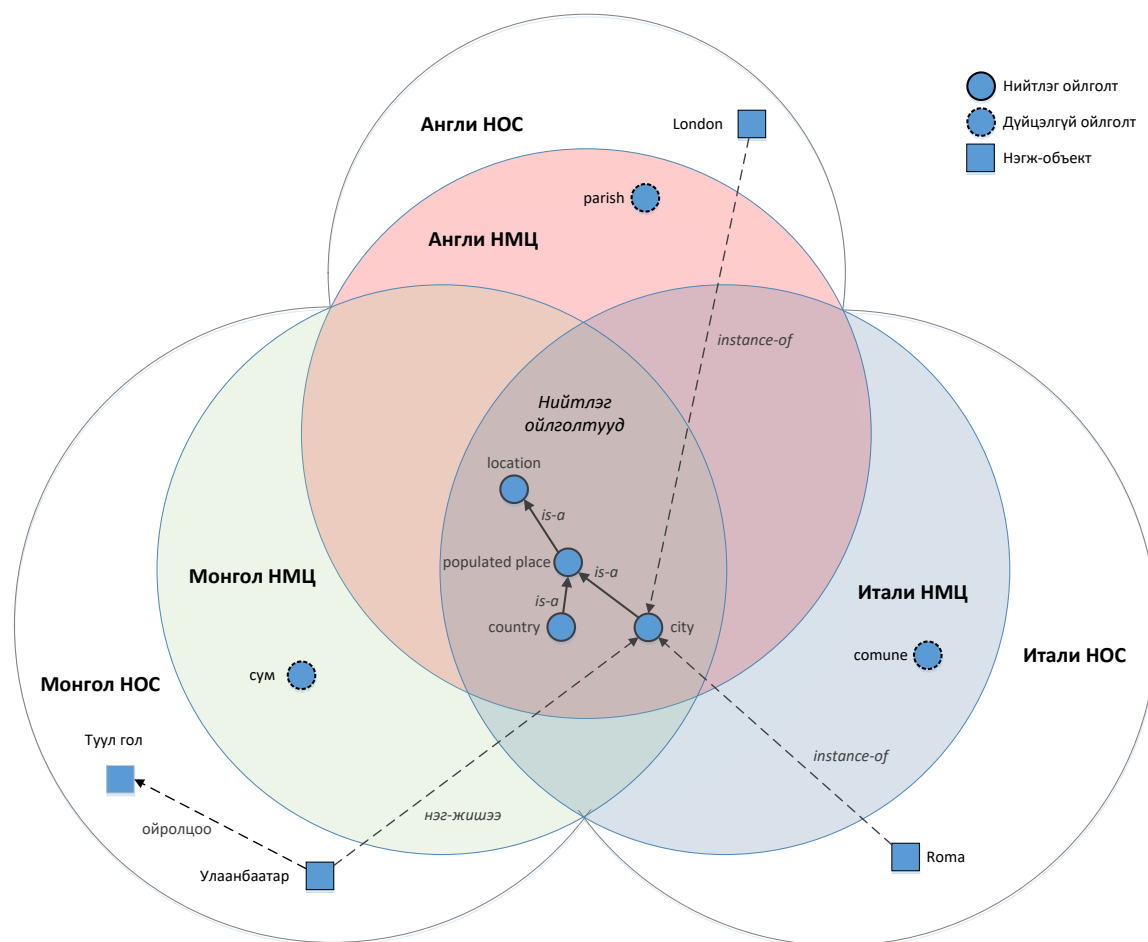
1.4 Ялгамжийн асуудал

Энэ бүлэгт онтологийг үүсгэхэд тавигдах ялгамжийн асуудлыг тодорхойлох болно. Ялгамж (diversity) гэдгийг ерөнхий утгаар нь аливаа зүйлсийн ялгаатай байдал эсвэл ялгаатай үзэгдэл гэж ойлгож болно. Мэдлэгийн сан, семантик технологийн хувьд ялгамж нь өгөгдлийн түвшинд, хэл болон үгийн сангийн түвшинд болон формаль онтологийн түвшинд илэрч болно. Энэ диссертацийн ажлаар ЕМС-ийн хүрээнд ялгамжийг ойлголтын болон хэлний цөмийн түвшинд авч үзлээ. Зураг 1.10-т итали, англи, монгол хэл соёлын хооронд зэрэгцэн орших нийтлэг ойлголтууд болон тэдгээрийн ялгамжийг ойлголтын түвшинд үзүүлэв. Энд *city, country, populated place* зэрэг ойлголтууд нь энэ гурван соёл дунд ижил утгаар хэрэглэгддэг байхад *сум, parish, comune* зэрэг ойлголтууд зөвхөн тухайн соёлд л хэрэглэгддэг. Сум гэдэг ойлголт Монгол улсын засаг захиргааны нэгжийн бүтцэд хэрэглэгддэг ойлголт бөгөөд энэ ойлголттой шууд утгаар дүйх ойлголт бусад хэл соёлд байдаггүй. Ийм учраас ЕМС эдгээр нийтлэг болон ялгаатай ойлголтуудыг бүгдийг нь агуулсан нэгдмэл мэдлэгийн сан болно.

1.4.1 Олон хэлний ялгамж

Нэгэнт олон хэл соёл өөр хоорондоо ялгарах маш олон ойлголтуудыг тэдгээрийн ахуй, ёс заншил, засаг захиргаанаасаа хамааруулан хэрэглэдэг тул олон хэлний

ялгамжийн асуудал үүснэ. Хэл гэдэг бол тухайн орон нутгийн ард иргэдийн харилцааны хэрэгсэл бөгөөд хэл болгон өөрийн гэсэн үгийн сантай байдаг. Үгийн сангийн түвшинд үг, түүний утгалбар, үглэврийн ялгаатай байдал байнга тохиолдоно. Онтологи бол аливаа хэл соёлоос салшгүй холбоотой, байгаль дээрх ойлголтуудыг үгийн сангаар илэрхийлдэг тухай өмнөх дэд бүлгүүдээр тайлбарласан билээ. Иймд бид олон хэлний ялгамжийг хэлний цөмийн талаас зөвхөн үгийн сангийн түвшинд авч үзэх юм.



Зураг 1.10 Ертөнцийн мэдлэгийн сангийн ялгамж

Ялгамж нь хэлний цөмийн дараах элементүүдэд ихэвчлэн тохиолддог. Эдгээрээс жишээ авч үзье.

1. Үглэвэр

Аливаа хэлний үглэвэр бичгийн системийн дагуу бүтдэг бөгөөд хоёр өөр бичгийн систем ашигладаг хэлнүүдийн хувьд ялгаатай байх нь ойлгомжтой. Гэвч үглэврийн ялгамж нь ижил бичгийн систем хэрэглэдэг нэг хэлний хувьд бас тохиолдоно. Жишээ нь, 1.3.3 дугаар бүлэгт үзүүлсэн жишээгээр испани хэлний *alojamiento* (Испани улсад),

эсвэл *hospedaje* (Өмнөд Америкт) зэрэг үгс нь нэг утгаараа нэг ойлголтыг илэрхийлж байгаа. Үүнтэй төстэй жишээг британи англи хэл, америк англи хэлний хувьд олныг нэрлэж болно.

2. Утгалбар

Нэг үг олон утгалбартай байдаг. Хэдийгээр тухайн хэлний үг болгон тодорхой тооны утгалбараараа ялгарч байхад нийтлэг утгалбаруудтай үгс бас байна. Энэ нь хоёр өөр хэлний үгс 1:1 харьцаатай харилцан орчуулагдах бөгөөд тэдгээр үгс утгалбарын түвшинд ч ижил байх тохиолдол юм. Жишээ нь, англи хэлний *head*, монголоор *толгой* гэх үгсийн утгалбар 1) *хүн амьтны биеийн эрхтэн*, 2) *тэргүүн*, *толгойлогч* гэж хоёр янз байна [43].

3. Синсет

Нэгэнт 1:1 орчуулга байгаа тохиолдолд синсенийн үгийн тоогоороо ижил байж болно. Жишээ нь, англи хэлний нэг үгтэй синсет *{city}* нь монгол хэлний нэг үгтэй *{хот}* гэсэн синсеттэй дүйх юм. Ихэнх тохиолдолд нэг ойлголтыг хоёр өөр хэлний синсет үгийн тоогоороо харилцан адилгүй байх нь олонтаа.

Энэ мэтчилэн үгийн үглэвэр болон утгалбар, синсенийн түвшинд төрөл бүрийн ялгамж байж болно. Нэгэнт үгийн утгалбар, синсенийн түвшинд ялгамж байгаа үед тэдгээрийг холбосон үгзүйн болоод үг-утгазүйн холбоосууд ялгаатай байх нь ойлгомжтой.

Энэ бүгдээс гадна аливаа хэлний үгийг зорилтот хэлэнд огт орчуулах боломжгүй бол энэ нь ойлголтын ялгамж буюу дүйцэлгүй ойлголтыг бий болгоно.

1.4.2 Дүйцэлгүй ойлголт

Хэрэв нэг ойлголтыг тухайн хэл дээр үгийн сангийн нэгжээр илэрхийлж чадаж байхад өөр хэл дээр үгсийн чөлөөт хувилбараар (free combination of words) илэрхийлж байвал үүнийг дүйцэлгүй үг (lexical gap) гэнэ. Дүйцэлгүй үг үргэлж дүйцэлгүй ойлголтыг илэрхийлж байдаг. Жишээ нь, англи хэлний *parish* (*Английн засаг захиргааны 3-р түвшний нэгж*) гэдэг үг нь Монгол хэлэнд орчуулагдах боломжгүй, өөрөөр хэлбэл, монгол хэлний үгийн санд энэ үгийн утгалбартай дүйх утгалбартай үг байдаггүй гэсэн санаа юм. Иймд энэ *parish* гэх ойлголт монгол хэлэнд дүйцэлгүй ойлголт болно.

Эрдэмтэн Л.Бентивойли [47] англи хэлнээс итали хэлний дүйцэлгүй үгийг 40 мянган толгой үгтэй, 60 мянган утгалбар бүхий Англи-Итали толь бичгээс шүүж үзэхэд 7.8% нь дүйцэлгүй үг байсныг илрүүлсэн байна. Тэрээр дүйцэлгүй үгийг тодорхойлохдоо үгийн сангийн нэгж гэдэгт толь бичгийн толгой үг, өвөрмөц хэллэг (idiom), хэвшмэл хэллэг (restricted collocation) зэргийг нэрлэсэн байна. Толь бичгийн толгой үгийг үгийн сангийн нэгж гэж шууд үзэж болох бол өвөрмөц болон хэвшмэл хэллэгийг нэг утга илэрхийлдэг гэдэг талаас нь авч үзжээ.

Өвөрмөц хэллэг бол хэд хэдэн үгээс тогтсон нийлэмж бөгөөд бүхэлдээ нэг утга илэрхийлэх нийлэмж [43]. Бүрдүүлбэр үгийг ойролцоо утгатай үгсээр сольж болохгүй. *Хэвшмэл хэллэг* бол хэд хэдэн үг нийлж хэвшсэн утгыг илэрхийлэх бөгөөд бүрдүүлбэр үгийг хязгаарлагдмал хүрээнд сольж болдог нийлэмж юм. Үүнийг бусад хэлэнд үгчлэн орчуулах боломжгүй байдаг [47]. Ийм дан ганц ойлголтыг илэрхийлэх өвөрмөц болон хэвшмэл хэллэг, толь бичгийн толгой үгээс өөрөөр буюу үгсийн чөлөөт хувилбараар илэрхийлэгдэх үгсийг дүйцэлгүй үг гэж үзжээ. *Үгсийн чөлөөт хувилбар* нь өгүүлбэрзүйн энгийн дүрмийн дагуу чөлөөтэй бүрэлдэн тогтсон үгсийн нийлэмж юм.

Дүйцэлгүй үгийг аль онтологид дүйцэхгүй байгаагаас нь хамаарч эх онтологи дахь дүйцэлгүй үг, зорилгын онтологи дахь дүйцэлгүй үг гэж хоёр салган нэрлэж болно. Онтологи нутагшуулах явцад эх онтологийн дүйцэлгүй үгийг зорилгын онтологид хялбархан тэмдэглэн үлдээж болно. Харин зорилгын онтологийн дүйцэлгүй үгийг хэрхэн тодорхойлох нь тодорхойгүй асуудал юм. Өөрөөр хэлбэл, дүйцэлгүй үгийн илэрхийлэх ойлголт нь ЕМС-д шинэ ойлголт болно. Жишээ нь, монгол соёлд байх *сум*, *гэрийн буурь* зэрэг ойлголт англи соёлд дүйцэлгүй ойлголт болно. Харин ийм эх онтологийн дүйцэлгүй ойлголтыг олж тодорхойлон ЕМС-д нэгтгэн хадгалах нь шийдэгдээгүй нэг асуудал юм.

Бүлгийн дүгнэлт

Энэ бүгдээс харахад онтологи нь семантик вэб технологид зайлшгүй шаардлагатай, үндсэн логикийг илэрхийлэх бүрдэл юм. Онтологийг хэрэгжүүлсэн хэл (хүний хэл эсвэл машины хэл), агуулгын хамрах хүрээнээс хамаарч бүлэглэн дотор нь олон ангилдаг. Бас онтологийн ойлголтууд, тэдгээрийн хоорондын утгазүйн холбоосыг байгуулсан хэлбэр, онтологийн ерөнхий бүтэц талаас ангилал болон дүрслэл онтологи

гэж ангилдаг. Дүрслэл онтологийн хувьд өнөөгийн семантик технологид илүү нийцтэй, мэдлэгийг боловсруулахад өргөн боломжтой гэдэг нь харагдаж байна.

Ийм олон төрөл зүйлийн онтологийг нутагшуулах аргачлалыг эрдэмтэд олон янзын хувилбараар тодорхойлсон байна. Энэ судалгааны ажлын хүрээнд судалсан аргачлалууд ихэвчлэн хагас автомат аргад түшиглэсэн бөгөөд хэлний нөөц, хэл боловсруулалтын хэрэгсэл програм ашигладаг учраас хэлний нөөц дутмаг, хэл боловсруулалтын технологи сайн хөгжөөгүй хэлнүүдийн хувьд ашиглах боломжгүй байна. Мөн онтологийг нутагшуулах давхарга, олон хэлний онтологи байгуулахад тулгарах асуудлууд, онтологийг нутагшуулалтын хэмжүүр зэргийг тодорхой түвшинд судалж үр дүнд хүрсэн боловч ялгамжийг хэлний болон ойлголтын хүрээнд нэгэн зэрэг тусгах асуудлыг судлах шаардлагатай байна.

ЕМС-ийн зохион байгуулалт, ялангуяа ойлголтын болон хэлний цөм нь онтологийн үгийн сангийн болон ойлгомжийн давхаргыг илэрхийлэх агаад эдгээр бүрдлийг онтологи талаас харвал онтологийн олон ангиллыг өөртөө багтаасан, онтологи нутагшуулалтын олон төрлийг шаардсан, олон хэлний дүрслэл онтологи гэдгийг харж болно. Ийм олон хэлний нэгдмэл онтологи үүсгэхийн тулд ялгамжийг тусгах хэрэгтэй. Нөгөө талаас онтологийг нутагшуулах цаг хугацаа, хүч хөдөлмөр их шаарддаг тул үүнийг хөнгөвчилсөн аргачлалыг боловсруулах шаардлагатай байна. Ялангуяа хязгаарлагдмал нөөцтэй хэл соёлын хувьд олны хүчээр богино хугацаанд, бага зардлаар онтологийг нутагшуулах замаар нэгдмэл онтологи үүсгэх боломжтой гэж үзэж байна.

БҮЛЭГ 2. ОЛНЫ ХҮЧИЙГ АШИГЛАХ СУДАЛГАА

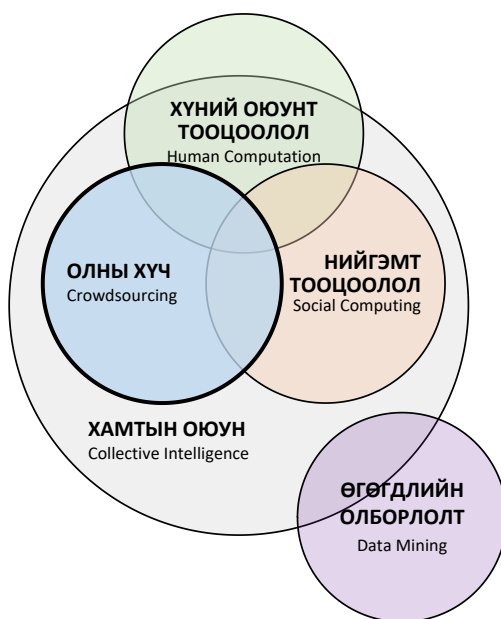
Энэ бүлэгт олны хүч, олны хүчийг орчуулгад хэрхэн ашиглах, орчуулгын чанарыг үнэлэх аргачлал, онтологийг нутагшуулах ажилд олны хүчийг ашиглах тухай тайлбарлана.

2.1 Олны хүч

Олны хүч гэдгийг ерөнхий утгаар тайлбарлавал олон нийтийн нэг хүн бүрийн хувь нэмрийг их хэмжээгээр цуглуулж хүний оюуны ашгийг хүртэх үйл явц гэж хэлж болно. Энэ нэр томъёог 2006 онд Жэфф Хау дараах байдлаар тодорхойлсон [48].

“Олны хүч гэдэг бол уламжлалт аргаар ажилчдад даалгавар өгч хийлгэдэг аливаа ажлыг маш олон хүнээр, нээлттэй байдлаар аутсорсинг хийлгэх үйлдэл юм”

Олны хүч бусад судалгааны чиглэл болох хүний оюунт тооцоолол (Human computation), нийгмийн тооцоолол (Social Computing) зэрэг судалгааны салбартай зарим талаараа давхцдаг (Зураг 2.1).



Зураг 2.1 Олны хүч ба бусад судалгааны чиглэл [49]

Хүний оюунт тооцоолол бол компьютерийн тооцооллоор шийдэж чадахгүй асуудлыг хүний оюуныг ашиглан шийдэх арга замыг судалдаг салбар. Олны хүч бол олон хүмүүсийн оюуны оролцоог нэгтгэж компьютерийн тооцооллоор хийж чадахгүй асуудлыг судалдаг салбар юм. Нийгмийн тооцоолол нь компьютерийн технологийн тусламжтайгаар хүмүүсийг нийгмийн харилцаанд оруулж өөр хооронд нь мэдээллийг

дамжуулснаар хамтын ажиллагааны үр дүнд асуудлыг шийдвэрлэх арга замыг судалдаг салбар болно.

Олны хүчний үндсэн арга техник бол аливаа ажлыг олон жижиг даалгавруудад (task) хувааж, нэг даалгаврыг олон хүнээр хийлгэж дараа нь үр дүнг нэгтгэн гаргаж авахад оршдог. Ийм загвараар аливаа ажлыг гүйцэтгэх нь “олон хүн хамтарч ажилласнаар цөөн тооны хүнээс илүү ухаалаг шийдвэрийг гаргана”, “олон тооны хүнтэй бүлгэм нь өөр доторх хамгийн ухаалаг хүнээс илүү ухаалаг байдаг” гэсэн үзэл баримтлалыг дагаж буй хэрэг юм. Үүнийг олон нийтийн цэцэн ухаан (wisdom of the crowds) гэж тодорхойлжээ [50]. Иймд олны хүчээр хийлгэх ажил нь машинаар бус хүний оюунаар хийлгэх ажил байдаг. Үүнийг хүний оюуны даалгавар (ХОД - НТ - Human Intelligence Task) гэнэ. Энэ даалгаварыг гүйцэтгэдэг хүмүүсийг оролцогч¹⁵ гэнэ.

Хүснэгт 2.1 Олны хүч ашигласан системийн төрлүүд [51]

Хамтрах байдал	Системийн архитектур	Оролцогчдыг зохион байгуулах	Хийх ажил	Жишээ	Шийдэх асуудал	Тайлбар
Шууд	Бие даасан	Тийм	Үнэлэх - шүүмж, санал, тэмдэглэгээ	Amazon.com дахь санал өгөх, шүүмж	Олон тооны зүйлсийг үнэлэх (бүтээгдэхүүн, хэрэглэгч г.м.)	Хүмүүс бол олон талын шүүмжлэгч
			Хуваалцах - зүйлс - бичвэрэн мэдлэг - бүтэцлэгдсэн мэдлэг	- YouTube, Flickr - Yahoo! Answers, Quora - Google Fusion Tables, Piazza, Galaxy zoo	Олон тооны зүйлсийг бий болгож (тархсан эсвэл төвлөрсөн) хүмүүсийн дунд хуваалцах	Хүмүүс бол агуулга үүсгэгч
			Сүлжилдэх	- LinkedIn, Facebook	Нийгмийн сүлжээ байгуулах	Хүмүүс бол өөрсдөө бүрдэл
			Бүтээгдэхүүн бий болгох - програм хангамж - бичвэрэн мэдлэгийн сан - бүтэцлэгдсэн мэдлэгийн сан - систем - бусад	- Linux, Apache, Hadoop - Wikipedia, openmind - DBPedia, YAGO-NAGA - Freebase, Mahalo - Digg.com, Second Life	Бодит зүйлсийг бүтээх	Хүмүүс янз бүрийн үүрэгтэй
			Даалгавар гүйцэтгэх	- Хүмүүсийг олох, агуулга үүсгэх, орчуулга	Дурын асуудал	Хүмүүс янз бүрийн үүрэгтэй
Шууд бус	Бие даасан	Тийм	- Зорилтот тоглоом тоглох - Бооцоо тавих - Хувийн эрх ашиглах - Капча шийдэх - авах/зарах/дуудлага	- ESP - intrade.com - IMDB эрх - recaptcha.net - eBay, World of Warcraft	- зураг тэмдэглэх - үйл явдлыг таамаглах - кино дүгнэх - бичвэрийг тоон болгох - хэрэглэгчдийн бүлгэм байгуулах	Хүмүүс янз бүрийн үүрэгтэй

¹⁵ Оролцогч (worker) нь олны хүчээр гүйцэтгэх ажлыг хийх олон нийтийн нэг бие хүн. Олны хүчийг ашигласан системийн талаас оролцогч нь системийн хэрэглэгч болдог. MTurk платформын хувьд оролцогчийг англиар turker ч гэж нэрлэдэг.

		худалдаа, олон тооны тоглогчтой тоглоом тоглох			
Өөр системд суурилж ажиллах	Үгүй	- түлхүүрэн хайлтын систем - бүтээгдэхүүн худалдан авах - вэб сайт үзэх	- Google, Yahoo, Bing - Amazon-ны онцлог зөвлөмж - Өөрчилж болдог вэб сайт (жич, Yahoo! Нүүр хуудас)	- зөв бичих дүрмийн алдаа засах - бүтээгдэхүүний санал болгох - вэб сайтыг ашиглахад хялбар байдлаар зохион байгуулах	Хүмүүс янз бүрийн үүрэгтэй

Ийм арга техникийг ашигласан системийн төрлүүдийг Хүснэгт 2.1-д үзүүлэв. Энд шууд хамтарч оролцдог систем нь оролцогчид үнэлгээ өгөх, ямар нэг зүйлсийг хуваалцах, сүлжээ байгуулах, хамтарч бүтээгдэхүүн хөгжүүлэх болон төрөл бүрийн даалгаврыг гүйцэтгэх боломжийг олгодог. Өөрөөр хэлбэл хүмүүс ил оролцоотой. Харин шууд бусаар хамтарч оролцдог системүүд нь хүмүүс системийг хэрэглэх явцдаа шууд бус байдлаар хамтарч асуудлыг шийдвэрлэх юм. Жишээ нь, ESP тоглоомд хэрэглэгчид үгийн сүлжээ тоглох ба нэг хэрэглэгч тусдаа таасан үгсийг нэгтгэж үр дүнг гаргаж авдаг.

Олны хүчийг ашигласан системүүд 4 үндсэн асуудал [51]: 1) *оролцогчдыг хэрхэн зохион байгуулах, авч үлдэх*, 2) *оролцогчид юу хийх*, 3) *оролцогчдын гүйцэтгэсэн ажлыг хэрхэн нэгтгэх*, 4) *оролцогчдыг хэрхэн үнэлэх* байдаг.

2.1.1 Оролцогчдын зохион байгуулалт

Оролцогчдыг зохион байгуулахад өргөн хэрэглэдэг 5 арга бий.

1. Оролцогчдыг ажилд оролцохыг шаардах

Хэрэв шаардлага тавих эрхтэй бол оролцогчдыг тухайн ажилд оролцохыг шаардаж болно. Жишээ нь, компанийн захирал бүх ажилчиддаа (оролцогчид) даалгавар өгөх гэх мэт.

2. Оролцогчид хөлс төлөх

Amazon Mechanical Turk¹⁶ (заримдаа MTurk гэдэг) олны хүчийг ашиглах үйлчилгээг хувь хүн, байгууллагууд жижиг даалгаврууд байршуулдаг ба даалгаврыг гүйцэтгэсэн оролцогчдод бага хэмжээний хөлс төлдөг.

3. Сайн дурын оролцогчдыг татах

Энэ арга төлбөргүй, хэрэгжүүлэхэд хялбар байдаг тул хамгийн түгээмэл байдаг. Жишээ нь, Википедиа, Youtube зэрэг системүүд энэ аргыг хэрэглэдэг.

¹⁶ <https://www.mturk.com>

Энэ аргын нэг сул тал бол тухайн ажилд хэчнээн хүн оролцохыг таамаглахад хэцүү байдаг.

4. Хэрэглэгчид төлбөр төлөхийг шаардах

Энэ нь А системийн хэрэглэгчид, А системийн үйлчилгээний төлбөрөө төлөхийн оронд олны хүчийг ашигласан Б системд оролцох арга юм. Жишээ нь, вэб сайтад сэтгэгдэл үлдээхийн тулд капча (captcha) оруулна. Ингэж тухайн хэрэглэгчийг спам, вэб роботоос ялгаж хүн мөн эсэхийг нь тогтоодог. Энэ нь OCR (Optical Character Recognition) програмын таньж чадахгүй байгаа зурган бичвэрийг таниулахын тулд хүний туслалцаа авч машин сургалтын алгоритмаа сайжруулах, бичвэрийн тоон болгох гол санааг агуулдаг. Бас фото зураг дээр ямар объектууд байгааг асуудаг Гүүглийн системийг нэрлэж болно.

5. Хэрэглэгчийн ул мөрийг ашиглах

Энэ арга бол олон тооны хэрэглэгчтэй тогтвортой ажилладаг системийн хэрэглэгчийн оруулсан мэдээлэл, тэдний хийсэн үйлдлийг ашигладаг. Жишээ нь, хайлтын системд хайсан түлхүүр үгсээс зөв бичих үсгийн дүрмийн алдаа шалгах системийг хөгжүүлж болно. Гэвч тэдгээр хэрэглэгчдийн ул мөр аливаа олны хүчийг ашиглан шийдвэрлэх зорилгод хэр нийцэхийг тогтооход хэцүү байдаг.

Оролцогчдыг дээрх аргуудаар зохион байгуулсны дараа оролцогчдыг хэрхэн авч үлдэх, цаашдын тогтвортой ажиллагааг хангах арга техник бас чухал асуудал болно. Үүнд дараах аргуудыг хэрэглэж болно.

Хүснэгт 2.2 Оролцогчдыг авч үлдэх арга

1	Баярлуулах	Хэрэглэгчийн оруулсан хувь нэмэр хэрхэн өөрчлөлтийг бий болгож байгаа хэрэглэгчид даруй харуулах
2	Зугаатай байлгах эсвэл хэрэгтэй үйлчилгээ өгөх	Тоглож байхдаа хувь нэмрээ оруулах
3	Нэр алдарт өрсөлдүүлэх	Нэр алдарт хүрэх, үнэлүүлэхийн тулд тогтвортой оролцох
4	Уралдуулах	Уралдаан зохион байгуулж тэргүүлэгчдийг зарлах
5	Эзэмшүүлэх	Хэрэглэгч системийн зарим хэсгийг эзэмшиж байгаа гэсэн сэтгэгдлийг төрүүлснээр тэр хэсгээ улам хөгжүүлэхэд татан оролцуулах

Дээрх аргуудыг олон тооны оролцогчдыг авч үлдэхийн тулд оролцогчдыг зохион байгуулах аргуудтай янз бүрээр хослуулж хэрэглэдэг. Жишээ нь, эхний шатанд ажлын хөлс төлөх, оролцохыг шаардах аргаар оролцогчдыг цуглуулаад дараа нь сайн дурын зохион байгуулалт руу шилжүүлэх тохиолдол байж болно.

2.1.2 Оролцогчдын хийх ажил

Олны хүч ашигласан системүүд олон төрлийн ажлыг оролцогчдоор гүйцэтгүүлж болно. Жишээ нь, үнэлүүлэх, шүүмжлүүлэх, дүгнүүлэх, тэмдэглүүлэх; эсвэл хуваалцах, шинээр үүсгэх, сүлжих; эсвэл зураг дээрээс зүйлсийг олох, нягтлан шалгах зэрэг энгийн ажлуудыг хийлгэж болно. Үүнээс гадна, маш нарийн төвөгтэй ажлуудыг хийлгэх боломжтой. Жишээ нь, мэдлэгийн сан байгуулахад оролцогчид ойлголтыг тодорхойлох, нэгж-объектыг нэмэх, түүний шинжийг засварлах, үнэн худлыг тэмдэглэх, цаашилбал, логик дүрмүүд бичих, зөрчилдөөнтэй логикийг засах зэрэг бүр төвөгтэй ажлуудыг ч гүйцэтгүүлж болно. Гол нь ямар ажлыг ямар хүрээнд хийх, том ажлыг хэрхэн хялбар ажлууд болгон хуваах асуудлыг голлож олны хүчинд тохирсон ажлыг загварчлах шаардлагатай болдог. Үүнийг хийхэд дараах нөхцөлүүдийг тооцох хэрэгтэй.

1. Мэдлэгийн шаардлага

Олны хүч ашигласан системүүд оролцогчдыг хэд хэдэн бүлэгт, жишээ нь, зочин, энгийн хэрэглэгч, хянан тохиолдуулагч, админ гэх мэт хувааж тус бүрд нь тохирох оролцооны загварыг боловсруулсан байдаг. Тухайлбал, бага үнэлэмжтэй оролцогчид (зочин, энгийн хэрэглэгч г.м.) үргэлж хялбар даалгавруудыг гүйцэтгэх сонирхолтой байдаг. Жишээ нь, энгийн асуултад хариулах, өгүүлбэр засах, буруу өгөгдлийг тэмдэглэх гэх мэт даалгавруудыг хийх хүсэлтэй байдаг. Хэрэв даалгаврын ачаалал их бол хийх хүсэлгүй байдаг. Харин өндөр үнэлэмжтэй оролцогчид (хянан тохиолдуулагч, админ г.м.) хэцүү даалгавруудыг гүйцэтгэх хүсэлтэй байдаг.

2. Үзүүлэх нөлөө

Ажилчдын оруулсан хувь нэмэр олны хүч ашигласан системд хэр зэрэг нөлөөлөхийг тооцох хэрэгтэй. Жишээ нь, мэдлэгийн санг олны хүчээр үүсгэж байх үед буруу өгөгдлийг тэмдэглэх нь мэдлэгийн санд олон газар хэрэглэх логик гаргалгааны дүрмийг өөрчилснөөс хамаагүй бага нөлөөтэй. Иймд их нөлөөтэй ажлуудыг өндөр үнэлэмжтэй хэрэглэгчдээр гүйцэтгүүлэх нь оновчтой.

3. Машины оролцоо

Хэрэв олны хүч ашигласан систем даалгаврыг гүйцэтгэхэд ямар нэг алгоритмыг ашигладаг бол машинд хялбар даалгавруудыг машинд, хүмүүст хялбар даалгаврыг нь оролцогчдод өгөх хэрэгтэй. Жишээ нь, зураг болон бичвэрэн тайлбараар хоёр бүтээгдэхүүнийг ижил мөн эсэхийг шалга гэвэл энэ нь мэдээж машинд хэцүү даалгавар юм. Харин энэ даалгавар хүнд хялбар санагдана. Иймд хийж чаддаг зүйлээрээ нэг нэгнийгээ гүйцээсэн загвараар ажиллах нь зүйтэй байдаг.

4. Хэрэглэгчийн интерфейс

Хэрэглэгчийн интерфейс нь оролцогчид хэрэглэхэд хялбар байх хэрэгтэй. Жишээ нь, мэдлэгийн санг байгуулах үед нэгж-объектын шинжийн утгыг оноох нь хялбар үйлдэл байхад нарийн төвөгтэй бүтэцтэй өгөгдлийг оруулах нь мэдлэгийн сангийн схемийн талаар ойлголттой байхыг оролцогчдоос шаардах тул илүү хүнд байдаг.

Энэ бүгдээс харахад олны хүчээр хийлгэх даалгавар бол энгийн оролцогчдод аль болох энгийн, машины оролцоотойгоор тэдний даалгаварт тусалдаг, хэрэглэхэд хялбар, ойлгомжтой интерфейстэй, нэг даалгавраар цөөн үйлдэл хийдэг байх нь чухал юм. Харин өндөр үнэлэмжтэй оролцогчдын хувьд арай төвөгтэй асуудлыг шийдүүлэхээр ажлын даалгаврыг зохиомжилж болно.

2.1.3 Ажлын нэгтгэл

Ихэнх олны хүч ашигласан системүүд тоон үнэлгээнд тулгуурлан оролцогчдын гүйцэтгэсэн ажлуудыг нэгтгэдэг. Гэвч ажлыг нэгтгэх арга нь хэрэглээнээсээ хамаарч өөр өөр байдаг. Жишээ нь, Википедиа хэрэглэгчиддээ засвар өөрчлөлтийг нэгтгэх боломжийг олгодог, ESP (Extra Sensory Perception) тоглоом нь хоёр оролцогчийн санал болгосон үг ижил байвал автоматаар нэгтгэдэг. Иймд ажлыг хэрхэн нэгтгэх гэхээсээ илүү гол асуудал нь оролцогчид ялгаатай үр дүн үзүүлбэл яах вэ гэдэг нь гол асуудал юм. Үүнийг автомат болон гар аргаар шийдвэрлэх боломжтой. Автомат аргын хувьд голдуу өөр оролцогчдоос тухайн ажилд үнэлгээ авч түүнд түшиглэн асуудлыг шийдвэрлэдэг. Жишээ нь, мэдлэгийн сан байгуулахад оролцогч бүрийн хувь нэмэр үнэн байх магадлалыг тооцох замаар шийдвэрлэж болно. Харин гар аргын хувьд оролцогчдыг хооронд нь тэмцэлдүүлж сүүлд нь зохицуулах замаар асуудлыг шийдвэрлэдэг. Жишээ нь, Википедиа энэ аргыг ашигладаг бөгөөд өгүүллэгийн маргаантай агуулгыг оролцогчдынхоо санал нийлэмж дээр түшиглэн шийдвэрлэдэг байна. Эндээс харахад автомат арга нь энгийн санал асуулга, тоон үнэлгээг ашиглан

хялбар шийддэг бол гар арга нь илүү төвөгтэй, машинаар боловсруулахад төвөгтэй ажлуудад ашигладаг байна.

2.1.4 Ажлын үнэлгээ

Олны хүчийг ашигласан систем ажлыг буруу гүйцэтгэдэг, шаардлага хангахгүй оролцогчдыг илрүүлж улмаар тэдний оролцоог хязгаарлах, хориглох шаардлагатай байдаг.

1. Оролцоог хязгаарлах

Шаардлага хангахгүй оролцогчдын оруулсан хувь нэмрийг хязгаарлаж болно. Жишээ нь, оролцогчдын оруулсан өгөгдлийг мэргэжилтнээр хянуулах, засварлуулсны дараа үр дүнг хүлээн авдаг байж болно.

2. Илрүүлэх

Оролцогчдын оруулсан хувь нэмрийг мэргэжилтнээр үнэлүүлж шаардлага хангахгүй оролцогчдыг илрүүлж болно. Бас оролцогчдыг сорилоор шалгаж болно юм. Жишээ нь, оролцогчдоос хариу нь тодорхой асуултуудыг тавьж шалгах маягаар оролцогчдын мэдлэг, хувь нэмрийн чанарыг тодорхойлж болдог.

3. Оролцоог хориглох

Шаардлага хангахгүй оролцогчдын хувь нэмрийг хүлээж авахгүй байж болно. Хамгийн энгийн арга бол олны хүч ашиглах үйл ажиллагаанд оролцуулахгүй байх юм.

Олны хүчийг ашиглах нь эдийн засаг болоод цаг хугацааны хувьд өндөр үр ашигтай [48], [50]–[54]. Иймд онтологийг нутагшуулахад олны хүчийг ашиглавал тохиромжтой байж болох юм. Учир нь олон хэлний онтологийг нутагшуулах явцад ялгамжийг өндөр чанартай тодорхойлоход хүний оюуныг ашиглах нь зүйтэй. Мөн энэ ажил цаг хугацаа, хүч хөдөлмөр шаарддаг тул асуудлыг олны хүчээр богино хугацаанд бага зардлаар шийдвэрлэх боломжтой. Нэгэнт онтологийг нутагшуулах ажилд онтологийн үгийн сангийн орчуулга үндсэн даалгавар тул бичвэр орчуулгын даалгаварт олны хүчийг ашигласан судалгааг авч үзэх шаардлагатай.

2.2 Орчуулга ба олны хүч

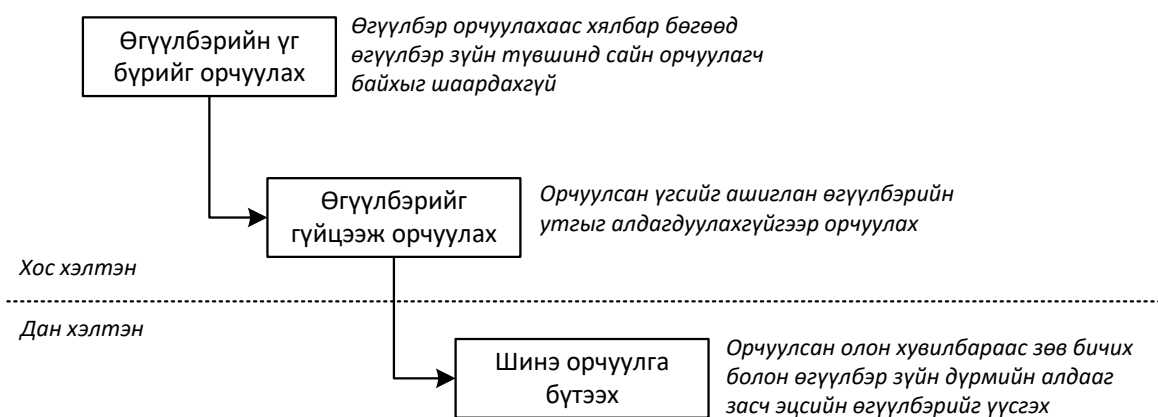
Аливаа орчуулгын ажил нь цаг хугацаа шаардсан нүсэр ажил байдаг бөгөөд чанартай үр дүнг мэргэжлийн орчуулагчаас авдаг. Харин статистик түшиглэсэн машин орчуулгын үр дүнг сайжруулахын тулд аль болох их хэмжээний өгүүлбэрүүд хадгалсан өндөр чанартай зэрэгцээ материалын сан маш чухал байдаг. Гэтэл машиныг

өндөр түвшинд сургах ажилд хүний оролцоо, мэргэжлийн орчуулагчийн орчуулгууд шаардлагатай байдаг [54]–[56]. Нэг талаас машины үр дүнг сайжруулахын тулд орчуулгыг олны хүчээр хийж өндөр чанартай орчуулгыг гаргаж авах судалгаа өргөжиж байна. Нөгөө талаас аливаа орчуулгын ажлын хурдан хугацаанд өндөр чанартай, мөн бага зардлаар гүйцэтгэхэд олны хүчийг ашиглах арга техникийг боловсруулах судалгааны ажлууд ч олноор хийгдэж байна.

2.2.1 Даалгаврын зохиомж

Хамтын ажиллагаат орчуулгын ажлын урсгалын зохиомж [52], номын сангийн каталог бүрдүүлэхэд орчуулга ашиглах [57], мэргэжлийн бус орчуулгын чанарын удирдлагын загвар болон орчуулгын чанарыг шалгах [56], [54] зэрэг ажил олны хүчээр орчуулга хийх, орчуулгын чанарыг үнэлэх асуудлыг хөндсөн байна.

Хамтын ажиллагаат орчуулга [52] нь эхлээд хос хэлтэй оролцогчид өгүүлбэрийн үг бүрийг орчуулаад дараа нь өгүүлбэрийг гүйцээж орчуулдаг ба эцэст нь зорилтот хэлний дан хэлтэй оролцогчид тухайн өгүүлбэрт тохирох хамгийн зөв орчуулгыг үнэлж тогтоодог. Энэ ажлаар өгүүлбэрийг бүхэлд нь нэг орчуулагчаар орчуулуулах бус өгүүлбэрийн бүрэлдэхүүн хэсэг тус бүрд нягт хамтын ажиллагаатай орчуулгыг хийсэн нь байна (Зураг 2.2).



Зураг 2.2 Хамтын ажиллагаат орчуулгын нэг загвар

Номын сангийн каталогийг бүрдүүлэхэд номын хавтасны нүүрийг зураг хэлбэрээр оролцогчдод өгч нүүрэнд байх номын зохиогч, гарчиг, хэвлэлийн газар зэргийг өөр хэл рүү орчуулах даалгаврыг олны хүчээр гүйцэтгэх ажлыг туршсан байна [57]. Үүнд номын нүүр ямар хэл дээр бичигдсэнийг тодорхойлох, бичвэрийг орчуулах, хамгийн сайн орчуулгыг сонгож шаардлагатай бол засах, зөв орчуулсан эсвэл хэлийг

зөв тодорхойлсон оролцогчдод нэмэлт урамшуулал олгох гэсэн хамтын ажиллагаат аргыг хэрэгжүүлсэн байна.

Олны хүчээр орчуулсан орчуулгыг гар аргаар үнэлж мэргэжлийн орчуулагч шиг өндөр чанартай орчуулга гарган авч болдог, үүнийг олны хүчний платформ ашиглан хямд зардлаар гүйцэтгэх боломжтойг нотлон харуулсан байна [54], [56]. Ингэхдээ нэг ХОД-т 10 өгүүлбэр өгч 4 өөр оролцогчдоор орчуулуулж авсан байна. Дараа нь хүлээж авсан орчуулгын хувилбаруудыг шаардлагатай бол засах даалгаврыг 5 өөр оролцогчдоор, түүнчлэн өгүүлбэрүүдийг оновчтой байдлаар нь эрэмбэлэх даалгаврыг 3 өөр оролцогчдоор гүйцэтгүүлсэн. Эцэст нь мэргэжлийн бус оролцогчдоор хийлгэсэн чанарын үнэлгээ мэргэжлийн орчуулагчийн түвшинд байна гэж дүгнэсэн. Энд бас нэг шинэ үнэлгээний арга туршсан нь чанарын удирдлагын загварт хүний уншиж ойлгох чадварыг ашигласан ба эхээс асуулт асууж зөв хариулж байгааг тооцоолжээ. Өөрөөр хэлбэл, орчуулсан бичвэрээс хүмүүс юуг ойлгож байгаагаар нь орчуулгын үнэн зөвийг үнэлсэн байна.

Дээрх аргууд олны хүчний орчуулгыг гар аргаар үнэлэх болон дэс дараатай олон давтагдах жижиг даалгавруудад хувааж гүйцэтгүүлэх, нэг даалгаварт дунджаар 5 орчим орчуулга хүлээн авч нэгтгэх маягаар ХОД боловсруулсан байна. Ажилчдын орчуулгын үр дүнг өмнө нь бэлдсэн мэргэжлийн жишиг орчуулгатай харьцуулах, мэргэжилтнүүдээр үнэлүүлэх, оролцогчдоор үнэлүүлэх чанарын үнэлгээний загвар боловсруулж үнэлснээр хямд зардлаар чанартай үр дүн гарган авч болохыг харуулжээ. Иймд ямар аргаар орчуулгын чанарыг үнэлсэн нь сонирхолтой байна.

2.2.2 Чанарын үнэлгээ

Олны хүчээр гүйцэтгэсэн орчуулгын чанарыг мэргэжилтнүүдийн орчуулсан жишиг орчуулгатай харьцуулж BLEU (Bilingual Evaluation Understudy) [52], [54], [56], HTER (Human-mediated Translation Edit Rate) [56], WER (Word Error Rate) [57] оноо бодох замаар үнэлдэг.

2.2.2.1 BLEU (Bilingual Evaluation Understudy)

BLEU [58] бол нэг хэлээр бичсэн бичвэрийг нөгөө хэл рүү машин орчуулгын аргаар орчуулж гарч ирэх бичвэрийн чанарыг үнэлдэг алгоритм. Гол зарчим нь машинаар орчуулсан орчуулга нь мэргэжлийн хүний гүйцэтгэсэн орчуулгатай аль болох ойр байх ёстойд оршино. Ямарваа орчуулгыг мэргэжлийн хүн үнэлж дүгнэлт гаргахад хамгийн ойр үр дүн гаргадаг хамгийн анхны үнэлгээнүүдийн аргуудын нэг бөгөөд

өнөө үед хамгийн түгээмэл хэрэглэгддэг, үр ашигтай аргуудын нэгт тооцогддог. Орчуулгын чанарыг тооцохдоо бичвэрийг сегментүүдэд (голдуу өгүүлбэрүүдэд) хувааж тус бүрийг нь мэргэжлийн жишиг орчуулгуудтай харьцуулах замаар BLEU оноо боддог. Дараа нь нийт орчуулгын өгүүлбэр бүрийн онооны дунджийг авч үзнэ. BLEU оноо үргэлж 0-ээс 1-ийн хооронд байна.

Энэ арга нь өөрчилсөн n-gram нарийвчлалын (modified precision) хэмжүүрийг ашигладаг тул цөөн үгтэй өгүүлбэрийг үнэлэх эсвэл харьцуулах жишиг орчуулга нь цөөн байхад хангалтгүй, өөрөөр хэлбэл 1 рүү дөхсөн оноо гарна. Учир нь цөөн үгтэй өгүүлбэр жишиг орчуулгад тохиолдох магадлал өндөр юм. Гол санаа нь өгүүлбэрт байгаа үгсийн n-gram тохирлыг жишиг орчуулгаас хайж хэр давхцаж байгааг тооцоолдог. Харин олон өгүүлбэртэй бичвэрийг харьцуулах жишиг орчуулга ихтэй материалын сантай харьцуулж үнэлэхэд илүү үр дүнтэй.

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1-\frac{r}{c})}, & \text{if } c \leq r \end{cases} \quad (2.1)$$

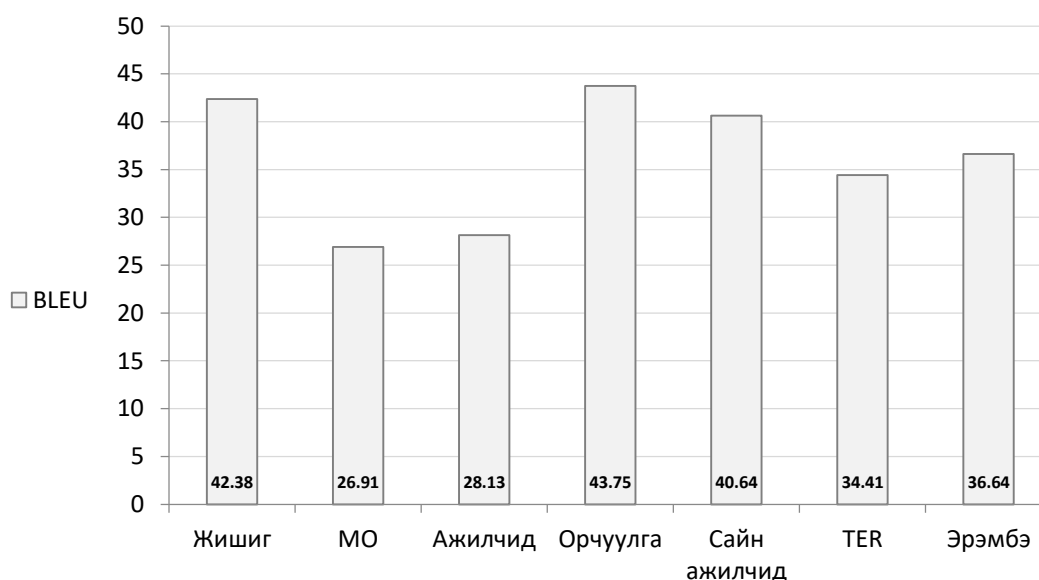
Хэрэв шалгах өгүүлбэрийн урт, c нь жишиг өгүүлбэрийн уртаас их байвал өөрчилсөн n-gram аргаар жишиг өгүүлбэрийн урт, r -д ойртуулж шалгах өгүүлбэрийн n-gram-ыг тасдаж тооцдог. Эсрэг тохиолдолд өндөр оноо авах шалгах өгүүлбэрийн урт жиших өгүүлбэрийн урттай тохирч байх ёстой. Үүнийг богинохон байх торгууль (brevity penalty) гэдэг (2.1) бөгөөд BLEU оноог (2.2) бодохдоо энэ утгаар үржүүлж өгнө [58].

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (2.2)$$

Энд p_n бол жишиг материалын сангийн N урттай өөрчилсөн n-gram-ын геометр дундаж, w_n бол эерэг жин $w_n = 1/N$. N -ийн урт нь 4 байхад тохиромжтой гэж үздэг.

О.Зайдан [54] нар олны хүчний орчуулгын чанарыг тооцох загварыг өгүүлбэрийн түвшний, оролцогчдын түвшний болон орчуулгын эрэмбийн түвшний нийт 21 онцлогийг тодорхойлж боловсруулсан байна. Өгүүлбэрийн түвшинд BLEU алгоритмд хэрэглэдэг n-gram тохирлын хувь, n-gram 3, 4, 5 байх үеийн тохирлын хувийн геометр дундаж, нэг орчуулгыг бусад хувилбаруудтай харьцангуй засварласан давтамжийн дундаж гэх мэт онцлогуудыг тодорхойлсон байна. Ажилчдын түвшинд орчуулгын түвшний онцлогуудын тухайн оролцогчийн хувьд тооцсон утга, ажиллаж эхлэхээс өмнө асуусан хэлний мэдлэг (унаган хэл, хэр удаан ярьж байгаа г.м.), оролцогчийн

амьдарч буй байрлал зэргийг тодорхойлсон. Эрэмбийн түвшинд тухайн орчуулгын дундаж эрэмбэ, хамгийн сайн эрэмбэ, сайн эрэмбэ зэрэг онцлогийг хамруулжээ. Энэ загварт тодорхойлсон онцлогийн түвшин бүрээр харьцуулсан туршилтууд хийсэн байна. Ингэхдээ мэргэжилтнүүдийн 4 жишиг орчуулгын (жишиг) 1-ийг нөгөө гуравтай нь харьцуулах замаар бодсон BLEU онооны дунджаар мэргэжлийн түвшний чанарыг тодорхойлсон байна. Улмаар олны хүчээр гүйцэтгэсэн орчуулга, нэг сайн машин орчуулгын системийн [59] үр дүн (МО - Машин орчуулга), TER (Translation Edit Rate), орчуулгын эрэмбэлсэн үр дүнтэй харьцуулжээ. Олны хүчээр гүйцэтгэсэн орчуулгын BLEU оноог бодохдоо 2 туршилт хийсэн. Нэгдүгээрт (Орчуулга), олны хүчээр орчуулсан 4 орчуулгаас жишиг орчуулгатай харьцангуй хамгийн өндөр оноотойг сонгосон байна. Хоёрдугаарт, (Сайн оролцогчид) хамгийн өндөр оноотой орчуулга хийсэн оролцогчдын гүйцэтгэсэн орчуулгаас өндөр оноотойг нь сонгож мэргэжлийн түвшний чанартай харьцуулсан байна. Мөн үнэлгээний зарчимд тулгуурлан 2 туршилт хийсэн. Эхнийх (TER) нь оролцогчдын гүйцэтгэсэн 4 орчуулгаас хамгийн бага TER дундажтай орчуулгыг, удаах (Эрэмбэ) нь хамгийн сайн дундаж эрэмбэтэй орчуулгыг (оролцогчид эрэмбэлсэн) сонгож харьцуулсан байна (Зураг 2.3). Мөн олны хүчний орчуулгын (Ажилчид) дунджийг бодсон байна.



Зураг 2.3 Олны хүчээр гүйцэтгэсэн орчуулгын BLEU оноог олон аргаар бодсон туршилтын үр дүн [54]

Дээрх зургаас харахад олны хүчээр гүйцэтгэсэн орчуулга нь мэргэжлийн түвшинд дөхөж очсон бөгөөд судлаач О.Зайдан нар олны хүчээр мэргэжлийн түвшний орчуулгыг гүйцэтгүүлэх боломжтойг нотлон харуулжээ. Мөн өгүүлбэр, оролцогчид

болон эрэмбийн гэсэн гурван түвшинд тодорхойлсон онцлогуудыг нэгтгэн түвшин тус бүрд BLEU оноог бодож үзэхэд харгалзан 34.71, 35.45, 37.14 байв.

2.2.2.2 HTER (Human-mediated Translation Edit Rate)

HTER [60] бол АНУ-ын батлах хамгаалах яамны DARPA агентлагийн GALE (Global Autonomous Language Exploitation) хөтөлбөрт бодит хугацаан дахь орчуулга хийхэд ашигладаг үнэлгээний арга юм. Энэ аргаар “засварлах зай” буюу машин орчуулгыг хэдэн удаа засварласны дараа мэргэжилтний орчуулгатай дүйцэхийг тодорхойлдог. TER (2.3) бол шалгах бичвэрт шаардлагатай өөрчлөлтийг хийхэд жишиг бичвэрийн аль нэгтэй тохирох хамгийн бага өөрчлөлтийн утгыг жишиг бичвэрийн дундаж уртад харьцуулсан харьцаа юм.

$$TER = \frac{\text{өөрчлөлтийн тоо}}{\text{жишиг бичвэрийн дундаж урт}} \quad (2.3)$$

Боломжит өөрчлөлтүүд нь үг оруулах, устгах, солих болон байрлалыг нь өөрчлөх зэрэг болно. Харин HTER бол хүн төвт буюу хүнээр засварлуулсан өөрчлөлтийг үнэлэх TER арга юм.

К.Каллисон-Берч [56] нар 5 машин орчуулгын системээр тус бүр 10 өгүүлбэртэй мэдээний бичвэрийг орчуулуулж дараа нь 5 өөр оролцогчдоос аль болох бага засвар оруулж хүний орчуулгын түвшинд өөрчлөх даалгавар өгсөн байна. Үр дүнд нь WMT09 [61] хурлаас гаргасан жишиг сантай харьцуулж олны хүчийг илүү төвөгтэй орчуулгыг үнэлэх HTER даалгавруудад ашиглаж болох ба олон орчуулгуудыг ч гаргах боломжтой гэж үзжээ.

2.2.2.3 WER (Word Error Rate)

Ж.Корний [57] нар орчуулгын чанарыг үнэлэхдээ яриа танилтын хувийг тооцдог WER-ийн өргөтгөсөн хувилбарыг (2.4) ашигласан байна.

$$WER_m = \frac{S + D + I + 0.5V}{N - T + K} \quad (2.4)$$

Энд S солигдсон үгийн тоо, D устсан үгийн тоо, I шинээр орсон үгийн тоо, N бол жишиг орчуулгад байгаа нийт үгийн тоо, V жишиг орчуулгад байгаа үгээс зөв бичих дүрмийн дагуу өөрөөр бичигдсэн үгийн тоо, T орчуулгад нэг үг хоёр үгээс бүтсэн нийлэмж болсон тоо, K орчуулгад хоёр үг болж задарсан үгийн тоо юм. Тэд энэ судалгаагаар урду хэлээр бичигдсэн номын хавтасны үгсийг англи хэл рүү орчуулах ажлыг Гүүгл машин орчуулгын системтэй харьцуулж WER_m оноог тооцсон боловч

Араб хэл дээр байгаа цөөн үгтэй орчуулгын хувьд олны хүчээр орчуулах нь машины чанараас харьцангуй бага байсан байна. Гэвч тэдний туршилт нь цөөн орчуулгын хувьд хийгдсэн нь үр дүнг илүү бодитой үнэлэхэд хангалтгүй байжээ.

2.2.2.4 Дотоод санал нийцэл

К.Каллисон-Берч [56] нар машин орчуулгын 5 системийн орчуулгыг, жишиг орчуулга болон эх бичвэрийн хамтаар 5 өөр оролцогчдод өгч сайнаас муу руу эрэмбэлэх даалгавар гүйцэтгүүлсэн байна. Дараа нь оролцогчдын үнэлгээ болон WMT08 хурлаас гарсан мэргэжилтнүүдийн үнэлгээ хооронд дотоод нийцэл (inter-annotator agreement) хэр байгааг тооцож үзсэн байна. Дотоод нийцлийг орчуулсан хос өгүүлбэр бүрийн хувьд хоёр үнэлэгч хоёул $A > B$, $A < B$, $A = B$ (өгүүлбэр) гэж үнэлэх тоог тоолж бодсон. Энд дотоод нийцлийн магадлал $1/3$ байна. Ажилчдын үнэлгээг авахдаа 5 өөр оролцогчдоор орчуулгуудыг эрэмбэлүүлж Шульцийн илүүд үзэж санал өгөх аргаар [62] нэг үнэлгээ болгон нэгтгэж авсан. Тэд мэргэжилтэн-мэргэжилтний хооронд дотоод санал нийлэмжийг тооцож үзэхэд 58%-д нь санал нийлсэн бол мэргэжилтнүүд зөвхөн 1 оролцогчтой санал нийлэх хувь нь 41% байсан байна. Харин оролцогчдын тоог 5 хүртэл өсгөхөд санал нийцлийн хувь өсөж байсан бөгөөд энэ тоо 5 байхад нийцлийн хувь 53% болж мэргэжилтэн-мэргэжилтний санал нийцлийн хувьтай ойртсон.

2.2.2.5 Корреляц

К.Каллисон-Берч [56] нар мөн судалгаагаараа оролцогчдын эрэмбийн үнэлгээ болон мэргэжилтнүүдийн эрэмбийн үнэлгээ хоорондын Спийрманы [63] корреляцыг тооцож үр дүн гаргажээ. Спийрманы корреляц (2.5) нь хоёр олонлогийн утгуудын эрэмбүүдийн хоорондох монотоник хамаарлын зэргийг тодорхойлдог. Өөрөөр хэлбэл, оролцогчдын эрэмбэлсэн утгууд мэргэжилтнүүдийн эрэмбэлсэн утгууд хооронд ямар хамааралтай байгааг гаргасан байна.

$$\rho_{rg_x} = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x} \sigma_{rg_y}} \quad (2.5)$$

Энд ρ эрэмбийн хувьсагчуудын Пийрсоны корреляцын коэффициент, $cov(rg_x, rg_y)$ эрэмбийн хувьсагчуудын коварианс, σ_{rg_x} болон σ_{rg_y} нь эрэмбийн утгуудын стандарт хазайлт болно.

Хэрэв Спийрманы корреляц нь эерэг бөгөөд 1-рүү дөхөх тусам оролцогчид болон мэргэжилтнүүдийн үнэлгээ хоорондоо илүү адилхан, нөгөө талаас оролцогчид

мэргэжилтнүүд шиг үнэлгээ өгч чаддаг гэсэн үг юм. Мэргэжилтэн-мэргэжилтэн хоорондын Спийрманы корреляц 0.78 байсан бөгөөд оролцогчдын үнэлгээг нь мэргэжилтнүүдийн үнэлгээтэй бараг ижил байсан байна.

О.Зайдан [54] нар Зураг 2.3-д үзүүлсэн аргуудад бодсон BLEU оноо болон жишиг орчуулгын BLEU оноо хооронд Пийрсоны корреляцын коэффициентийг бодож үзжээ. Энэ корреляц нь хоёр тоон олонлогийн хоорондох шугаман хамаарлын зэргийг тодорхойлно. Спийрманы корреляцаас ялгаатай шинж нь эрэмбийн утгуудыг тооцдоггүй бөгөөд олонлогийн утгууд дээр шууд бодолт хийдэг.

Энэ коэффициентын утгаас Зураг 2.3-д үр дүнтэй ижил сайн үр дүнг харсан байна. Орчуулга болон Сайн оролцогчид аргаар үнэлсэн BLEU оноонуудын коэффициентын утгууд харгалзан 0.81 (± 0.09 стандарт хазайлт), 0.79 (± 0.10) байсан нь жишиг орчуулгын коэффициенттой 0.81 (± 0.07) маш ижил байжээ. Харин TER 0.50 (± 0.26), Эрэмбэ 0.74 (± 0.17) корреляцтай байсан.

Энгийн орчуулгын чанарыг үнэлэхдээ мэргэжилтнүүдийн жишиг орчуулгатай харьцуулсан байна. Ингэхдээ орчуулгын чанарыг үнэлдэг BLEU, TER оноог тооцож улмаар мэргэжилтэн болон оролцогчдын үнэлгээ хооронд корреляцын хамаарлыг олох, дотоод санал нийцлийг тодорхойлох аргуудыг боловсруулсан байна. Мөн ХОД-ын зохиомжоос хамаарч дээрх аргуудыг боловсруулсан. Тэгвэл энгийн орчуулгаас гадна онтологийн үгийн санг орчуулах ажилд ХОД тодорхойлж түүнд тохирсон чанарын үнэлгээний аргыг боловсруулсан ажил бидний мэдэхээр судлаач Б.Лансер [22] нар байна. Энэ тухай дараагийн дэд бүлгээр өгүүлнэ.

2.3 Онтологи нутагшуулахад олны хүчийг ашиглах боломж

Хэдийгээр онтологийг нутагшуулахад орчуулгын ажил хийгдэх боловч энгийн орчуулга хийхээс өөр өөр ажлууд хийх шаардлагатай байдаг. Жишээ нь, онтологийн үгийн сан орчуулах, синсенийн үгийн утгалбарыг тодорхойлох, дүйцэлгүй ойлголтыг тэмдэглэх зэрэг байж болно.

Судлаач Б.Лансер [22] DBpedia мэдлэгийн сангаас жишээ болгон 10 элементийг сонгон авч түүнд хэрэглэгдэж буй үгийн санг 2 шатлалт ажлын урсгалтай даалгавар боловсруулж олны хүчээр англиас япон хэл рүү нутагшуулах замаар япон хэлний үгийн санг үүсгэх ажлыг гүйцэтгэсэн байна. Тэд үгийн санг үүсгэх даалгаврыг орчуулгын даалгавар буулгах санааг дэвшүүлж 1000 орчим үгтэй Англи үгийн санг гар аргаар үүсгэж бэлдсэн. Дараа нь олны хүчээр гүйцэтгэсэн орчуулгын чанарыг мөн

гар аргаар үүсгэсэн япон жишиг (gold standard) үгийн сантай харьцуулж маягаар үнэлсэн байна. DBpedia мэдлэгийн сангийн элементүүдийг нэрлэх үглэвэр нь олон хувилбаргүй, зөвхөн тогтсон нэг л үглэвэрээр илэрхийлэгддэг. Жишээ нь, *хүн* классын *гэр бүлийн хүн* гэдэг шинжийг англиар *spouse* гэдэг үгээр илэрхийлэх бөгөөд *wife of, husband of, to marry* гэх үглэврийн өөр хувилбарууд энэ шинжтэй холбоогүй байдаг байна. Онтологийн үгийн санг онтологийн бүтэц, семантик вэб болон онтологи инженерчлэлийн мэдлэг шаардахгүйгээр энгийн оролцогчдоор орчуулах, мөн нэг үг олон утгалбартай тул аль утгалбараар нь хэрэглэгдэж байгааг оролцогчдод зөв хэлж өгөхийн тулд DBpedia сангаас гурвалыг ашиглан орчуулах үглэврийг агуулсан өгүүлбэрүүдийг автоматаар [64] үүсгэж өгчээ. Ингээд эхний шатны даалгавар нь эдгээр автоматаар гарч ирсэн өгүүлбэрийг япон хэл рүү орчуулах, удаах даалгавар нь сайн орчуулгыг олж тэмдэглэх юм. Эхний даалгаврын хувьд элементийг илэрхийлэх үглэвэр бүрд 3 өгүүлбэр автоматаар үүсгэж 3 өөр оролцогчдоос орчуулахыг асуусан. Удаах даалгаврын хувьд нэг өгүүлбэрийн 3 орчуулга бүрийг бас 3 өөр оролцогчдоор асууж зөв орчуулгыг тэмдэглүүлсэн байна.

Мэхлэх хэрэглэгчдийг олохын тулд 20 шалгах асуулт, мөн тэдний ашигласан CrowdFlower¹⁷ платформуос санал болгож буй үнэнч хэрэглэгчийн баталгаа, газарзүйн байрлал (Японоос оролцож байгаа эсэх) зэргийг авч үзсэн байна. Эцэст нь энгийн дийлэнх олонхын саналаар зөв орчуулгыг тодруулж жишиг үгийн сантай харьцуулахад 3 саналаар тодруулсан орчуулгын f-score 0.78 байсан нь үүнээс бага буюу олон саналаар тодруулсан үр дүнгүүдээс хамгийн өндөр байсан. Үүний тохирол нь 0.87 байсан. Энэ судалгаагаар орчуулгыг дийлэнх олонхын саналаар үнэлүүлэхэд сайн чанартай үгийн санг үүсгэх боломжтой гэж үзсэн байна. Хэдийгээр тэд өгүүлбэр зүйн бүтэц, үгийн аймгийн тэмдэглэгээ, үгийн утгалбарын харьцангуй алдаа багатай өгүүлбэр автоматаар үүсгэж чадсан ч үгийн утгалбарын ялгааг олж харжээ. Жишээ нь, *yearOfConstruction* шинжийг илэрхийлэх үглэвэр нь япон хэл дээр *to completed in* гэж арай ерөнхий утгаар орчуулагдсан байна. Жишиг орчуулгад *constructed in* гэж илүү нарийвчилсан утгаар өгүүлбэрийг өгсөн ч япон оролцогчид ижил хариултыг өгчээ. Эндээс харахад хэл соёлын ялгаа, утгалбарын түвшинд ялгамжийн асуудал байна. Энэ ажилд [22] онтологийн үгийн сангаар өгүүлбэр бүтээж энгийн орчуулгад шилжүүлэн оролцогчдоор орчуулуулсан бөгөөд чанарын үнэлгээний арга нь өмнөх бүлэгт дурдсан аргуудтай ижил мэргэжлийн орчуулгатай харьцуулжээ. Тэд мэргэжлийн

¹⁷ <http://www.crowdfunder.com>

түвшний орчуулгыг гарган авах боломжтойг баталсан тул синсетийн тайлбар орчуулах зэрэг олон үгтэй бичвэр орчуулах ажлыг энэ аргуудын тусламжтайгаар орчуулж болно гэж үзэж байна. Тэгвэл үгсийг (синсетийг) орчуулахаас гадна зорилтот хэл дээрх синсетэд багтаж болох үгсийг нэмж оруулах тохиолдолд энэ бүлэгт дурдсан чанарын үнэлгээний аргуудыг судлаагүй байна. Бас үнэлгээ өгсөн оролцогчдын дотоод санал нийцлийг тооцож оролцогчдын хувь нэмрийг нэгтгэх аргууд сайн судлагдаагүй байна. Ийм учраас бид оролцогчдын оруулсан хувь нэмрийг үнэлэхдээ дотоод санал нийцлийн аргуудыг судалсан.

2.3.1 Олны санал нийцэл

Олны хүч ашигласан зарим системүүд оролцогчдын оруулсан хувь нэмрийг нэгтгэдэггүй эсвэл оролцогчдын гүйцэтгэсэн олон ажлуудаас энгийн олонхын саналаар сонгож нэгтгэдэг. Олон ажлыг нэгтгэж зөв хувилбарыг олохдоо оролцогчдоос үнэлгээ авч тэдгээрийн хоорондын санал нийцлийг тооцож үзэх ба хэрэв оролцогчид санал нийлж байвал тухайн ажлыг оролцогчид үнэн зөв үнэлсэн гэж үзэж болох юм. Тухайлбал, орчуулсан синсетийн үгийг зөв, буруу гэж оролцогчдоор үнэлүүлж болно.

Үнэлгээний үндсэн 2 төрөл байдаг:

- Чанарын (Qualitative)
 - Нэрлэсэн (Nominal): ямар нэг эрэмбэгүй эсвэл дараалалгүй утгуудыг үнэлэх;
 - жишээ нь, ангилах
- Хэмжих (Quantitative)
 - Дэс дарааллын (Ordinal): тодорхой дараалалд байх ч нарийн хэмжих боломжгүй утгуудыг үнэлэх;
 - Жишээ нь, аливаа зүйлийг тааламжгүй, бага зэрэг тааламжгүй, бага зэрэг тааламжтай, тааламжтай гэж үнэлэх
 - Завсрын (Interval): ямар нэг заах утгатай харьцуулж үнэлэх;
 - Жишээ нь, температур: усны хөлдөх болон буцлах температурыг зууд хуваасан завсрын утгуудыг үнэлэх
 - Харьцааны (Ratio): хэмжилт ба хэмжих нэгжийн харьцаа.
 - Жишээ нь, килограмм, метр гэх мэт хэмжих нэгжээр үнэлэх

Хэрэв олны хүчээр орчуулсан синсетийн үгийг зөв бурууг үнэлэх бол хэмжих үнэлгээний төрлийн аргуудыг ашиглах боломжгүй. Учир нь синсетийн үгс нь үнэн худлын хувьд ямар нэг дараалал, эрэмбэгүй утгууд болно. Жишээ нь, орчуулсан синсет 3 үгтэй болсон ба бүгд зөв байх тохиолдолд ямар нэг дараалалд орохгүй. Харин синсетийн эрэмбийг тогтооход дэс дарааны аргыг ашиглах боломжтой.

Үнэлгээний дээрх төрлүүдэд багтах статистикийн хэмжүүрүүдийг төрөл, санал өгөх оролцогчдын тоогоор Хүснэгт 2.3-д харьцуулан үзүүлэв.

Хүснэгт 2.3 Статистикийн хэмжүүрүүдийн ангилал

Үнэлгээний төрөл / санал өгөгчийн тоо	2 оролцогчид	2-оос олон оролцогчид
Чанарын аргууд (нэрлэсэн)	Cohen's kappa, Krippendorff's alpha	Fleiss' kappa, Krippendorff's alpha
Хэмжих аргууд (дэс дарааллын, завсрын, харьцааны)	Correlation, weighted Cohen's kappa, Krippendorff's alpha	Krippendorff's alpha, ICC (Inter-Class Correlation)

Дээрх хэмжүүрүүдээс корреляцыг орчуулгад хэрхэн ашигласныг тайлбарласан. Харин эдгээр хэмжүүрүүдээс онтологи нутагшуулалтад хэрэглэж болох хэмжүүрүүдийг судалсан.

2.3.1.1 Кохен каппа (Cohen's kappa)

Кохен каппа [65] нь ямарваа N зүйлийг C ангилалд (өөрөөр хэлбэл C тооны үнэлгээ өгөх) оруулан ангилж буй хоёр санал өгөгчийн хоорондын санал хэр нийцэж буйг хэмждэг статистик хэмжүүр (2.6) юм.

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (2.6)$$

Энд $Pr(a)$ нь санал өгөгч хоорондын харьцангуй нийцэл бөгөөд $Pr(e)$ нь үнэлгээнүүд тохиолдлоор нийцэх магадлал болно. Үүнд байгаа өгөгдлийг ашиглан үнэлгээ өгөгч бүр ямарваа ангилалд тохиолдлоор санал өгч болох магадлалыг олно. Үнэлгээ өгөгчдийн санал бүрэн нийцэж байвал $k=1$. Санал нийцэж буй байдал нь тохиолдлоор санал нийцэх магадлалаас хэтрэхгүй байвал $k=0$ гэж үзнэ. Утга -1 ээс 1 хооронд байдаг ба 0 – 0.20 байвал сул, 0.21 – 0.40 хооронд мэдэгдэхүйц, 0.41 – 0.60 бол нягт, 0.61 – 0.80 хүчтэй, 0.81 – 1 төгс нийцэл гэж үзнэ. Энэ каппа зөвхөн хоёр санал өгөгчийн

хоорондох нийцлийг хэмждэг. Хэрэв онтологийн үгийн санг нутагшуулахад үгийн орчуулгыг зөвхөн хоёр санал өгөгчөөр үнэлүүлнэ гэвэл энэ аргыг хэрэглэж болохоор байна. Өөрөөр хэлбэл орчуулсан синсенийн үгсийг олонлог гэж үзвэл хоёр санал өгөгч хоорондын нийцлийн зэргээр тухайн синсенийн орчуулга хэр зөв болсныг мэдэж болно. Эсвэл тэр синсенийн орчуулга санал зөрөлдөөнтэй буюу орчуулахад хэр төвөгтэй байгааг олж болох юм.

2.3.1.2 Флайс каппа (Fleiss' kappa)

Флайс каппа [66] бол хэд хэдэн зүйлийг ангилж (өөрөөр ангилан үнэлэх байж болно) буй хоёроос дээш тооны санал өгөгч байхад тэдгээрийн өгч буй үнэлгээнүүд хоорондоо хэр нийцэж байгааг шалгадаг статистик хэмжүүр юм. Ерөнхий зарчмын хувьд Cohen's kappa-гийн адилаар тохиолдлын байдлаар үнэлгээнүүд хоорондоо нийцэх магадлал гэсэн доод суурийг авч үзэж, энэ хэмжээнээс ямар илүү хэмжээгээр саналууд нийцэж байгааг тооцож гаргадаг.

Флайс каппа нэрлэн-харьцуулах эсвэл хоёртын байдлаар үнэлсэн олонлогууд дээр л ашиглах боломжтой. Эрэмбэлэх байдлаар ангилсан өгөгдөл дээр ашиглах боломжгүй байдаг. Нэрлэн-харьцуулах гэдэг нь ямар нэг зүйлсийг тэдгээрийн нэр эсвэл (мета-) ангилал болон бусад чанарын ангиллуудыг үндэслэн ялгах хэмжих утгын төрөл юм. Хоёртын төрөл нь үнэн/худал, тийм/үгүй, зөв/буруу гэх мэт ангилалтай байдаг.

Кохен каппад үнэлгээ өгөгчид нь хоёул ижил зүйлийн олонлогт үнэлгээ өгсөн тохиолдлыг авч үздэг бол Флайс каппад үнэлгээ өгөгчийн тоо нь тогтмол байх ба өөр өөр зүйлсэд өөр өөр хүмүүс үнэлгээ өгөх боломжийг олгодог. Өөрөөр хэлбэл нэг зүйлд A, B хүмүүс үнэлгээ өгөхөд өөр нэг зүйлд C, D гэсэн хүмүүс үнэлгээ өгч болно. Флайс каппаг томъёолбол (2.7):

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (2.7)$$

Үүнд $1 - \bar{P}_e$ нь тохиолдлын байдлаар санал нийлэхээс дээгүүр хэмжээнд санал нийлэх боломж, $\bar{P} - \bar{P}_e$ нь бодитой санал нийлсэн байдал болно. Хэрэв $k=1$ тохиолдолд санал бүрэн нийлсэн, $k \leq 0$ үед санал ер нийлээгүй гэж үзнэ (тохиолдлын байдлаар санал нийлэхээс илүү хэмжээнд хүрээгүй).

N тооны зүйлтэй, зүйл бүрд n тооны үнэлгээтэй (буюу n тооны санал өгөгч), k тооны ангилалтай байна гэж үзнэ. Зүйлсийг $i=1..N$ ширхэг, ангиллуудыг $j=1..k$ гэж үзье. n_{ij}

нь i -р зүйлийг j -р ангилалд оруулах санал өгөгчийн тоо эсвэл i -р элементийг j -р ангилалд оруулсан нийт санал гэж үзнэ. Эхлээд p_j буюу j -р ангилалд оноосон бүх саналын харьцааг олно. Энд i -р зүйлийн хувьд нийт саналын тоо n -тэй тэнцүү байх ёстой. Өөрөөр хэлбэл санал өгөгч бүр бүх зүйлсэд саналаа өгсөн байх ёстой.

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}, \quad 1 = \frac{1}{n} \sum_{j=1}^k n_{ij} \quad (2.8)$$

Дараа нь P_i буюу i -р зүйлийн хувьд санал нийлж буй харьцааг олно. Энэ нь тухайн зүйлд санал нийлсэн хосуудын нийт тоог санал нийлж болох бүх хосуудын тоо $n(n-1)/2$ -той харьцуулсан харьцаа юм. Харин i -р зүйлийг j ангилалд оруулах санал өгсөн нийт хос бол $n_{ij}(n_{ij}-1)/2$ юм. Эндээс P_i -ийг дараах томъёогоор бодож болно.

$$\begin{aligned} P_i &= \frac{1}{\frac{n(n-1)}{2}} \sum_{j=1}^k \frac{n_{ij}(n_{ij}-1)}{2} = \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij}-1) \\ &= \frac{1}{n(n-1)} \sum_{j=1}^k (n_{ij}^2 - n_{ij}) = \frac{1}{n(n-1)} \left[\left(\sum_{j=1}^k n_{ij}^2 \right) - (n) \right] \end{aligned} \quad (2.9)$$

Дараа нь P_i -уудын дундаж буюу \bar{P} -ийг (2.10) олно.

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^N \sum_{j=1}^k n_{ij}^2 - Nn \right) \quad (2.10)$$

Зарим тохиолдолд санал нийцэл зөв байж болох боловч хэрэв санал өгөгчид цэвэр санамсаргүйгээр санал өгвөл ангилал бүрийн хувьд санал нийцэх магадлалуудын дундаж харьцаа (2.11) томъёогоор бодно.

$$\bar{P}_e = \sum_{j=1}^k p_j^2 \quad (2.11)$$

Синсетийн орчуулгад оролцогчид хэр санал нийлж байгаа Флайс каппа ашиглан тооцож болох юм. Учир нь орчуулсан синсет бүр харилцан адилгүй тооны үгтэй байх ба санал өгөгчид нь өөр өөр синсетэд санал өгч болно. Нөгөө талаас зарим оролцогчид зарим синсетийн орчуулгыг үнэлэх, зарим нь өөрийг үнэлэх маягаар оролцогчдыг зохион байгуулах боломжтой. Хэрэв Кохен каппаг ашиглавал хоёр оролцогч л бүх

синсетийг үнэлэх шаардлагатай болно. Гэвч бидний даалгаварт нэг синсетийн орчуулга хэр үнэн болсныг үнэлж оновчтой үгсийг сонгох нь чухал юм.

2.3.1.3 Криппендорффийн альфа (Krippendorff's alpha)

Криппендорффийн альфа (α) нь ажиглагчид, санал болон үнэлгээ өгөгчид, үнэлэгчид аль эсвэл аливаа нэг таамаглагдашгүй үзэгдлийг хэмжих хэмжүүрүүд, эсхүл тэдэнд оноосон тооцож болохуйц утгуудын хооронд санал нийцлийг үнэлдэг үнэн зөвийн коэффициент юм [67]. Флайс каппа олонлогийн элементүүдийг хэдэн удаа аль ангилалд оруулсан тоон өгөгдлийг нэгтгэж тооцдог бол Криппендорффийн альфа нь олонлогийн аль элементийг хэн хэрхэн үнэлсэн тоон өгөгдлөөс санал нийцлийг бодно. Альфагийн коэффициентыг бодох томъёог (2.12)-д харуулав.

$$\alpha = 1 - \frac{D_o}{D_e} \quad (2.12)$$

Энд D_e бол хүрч болохуйц санал үл нийлэмж (2.13), харин D_o нь хүрчихээд байгаа санал үл нийлэмжийг (2.14) илэрхийлнэ.

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck \text{ metric}} \delta_{ck}^2 \quad (2.13)$$

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_{k \text{ metric}} \delta_{ck}^2 \quad (2.14)$$

Энд o_{ck} , n_c , n_k болон n нь тохиролцлын матрицаар тодорхойлдог утгууд болно. Хүрчихээд байгаа санал үл нийлэмж 0 байх тохиолдолд альфагийн утга 1 болж бүрэн санал нийлсэн гэж үзнэ. Хэрэв хүрчихээд байгаа санал үл нийлэмж хүрч болохуйц санал үл нийлэмждээ очвол альфа 0 болсноор санал нийлэмж байхгүй гэж ойлгоно. Альфа утга $0 \leq \alpha \leq 1$ хооронд байх ба үүнээс харвал ажиглалтаас үүссэн санал үл нийлэмж байж болох утгаасаа давж болдгийг сөрөг утга тооцогдох боломжоос харж болно. Альфагийн утгыг дараах жагсаалтад үзүүлснээр тайлж ойлгоно.

- $\alpha \geq 0.8$ – төгс санал нийцэл
- $0.8 > \alpha \geq 0.677$ – хүчтэй санал нийцэл
- $\alpha < 0.677$ – сул санал нийцэл

Криппендорффийн альфа дараах 4 өөр нөхцөлд ажилладаг.

1. Хоёртын өгөгдөлд хоёр хүн бүх элементэд санал өгсөн байх

2. Нэрийдсэн өгөгдөлд (nominal data) хоёр хүн бүх элементэд санал өгсөн байх
3. Нэрийдсэн өгөгдөлд хоёроос дээш хүн зарим элементэд санал өгсөн байж болох
4. Бүх төрлийн өгөгдөлд хоёроос дээш хүн зарим элементэд санал өгсөн байж болох

Зарим элементэд санал өгсөн байж болно гэдэг нь санал өгөгч тухайн элементэд санал өгөөгүй орхисон байж болно гэсэн үг.

Альфаг тооцоход санал өгсөн өгөгдлөөс үнэн өгөгдлийн матрицыг (reliability data matrix) (Хүснэгт 2.4) санал өгөгч, элемент гэсэн индексээр байгуулна.

Хүснэгт 2.4 Үнэн өгөгдлийн матриц

Үнэлэх элементүүд	1	2	...	u	...	N
<i>i-р санал өгөгч</i>	C_{i1}	C_{i2}	...	C_{iu}	...	C_{iN}
<i>j-р санал өгөгч</i>	C_{j1}	C_{ju}	...	C_{jN}
...
<i>u-д харгалзах санал өгөгчдийн тоо</i>	m_1	m_2	...	m_u	...	m_n

Элементүүд доторх тохиролцол гэдэг нь тухайн элементэд санал өгөгчийн саналуудаас хоорондоо таарах саналуудын хосыг олоод хос бүрийн хувьд нийт саналын нийлбэрийг харуулдаг (Хүснэгт 2.5).

Хүснэгт 2.5 Элементүүд доторх тохиролцлын хүснэгт

	<i>I</i>	.	<i>k</i>	.	
<i>I</i>	O_{I1}	.	O_{Ik}	.	n_I
.
<i>c</i>	O_{c1}	.	O_{ck}	.	$n_c = \sum_k o_{ck}$
.
	n_I	.	n_k	.	$n = \sum_c \sum_k n_{ck}$

Хэрэв бүх элементэд санал өгөөгүй буюу орхисон өгөгдөл байгаа бол O_{ck} нь $c-k$ гэсэн ангиллын хослолын тоонуудыг тухайн ангилалд нийт санал өгсөн хүмүүсийн тооноос нэгийг хасаж хуваасан утгуудын нийлбэр юм.

$$o_{ck} = \sum_u \frac{u \text{ элемент дэх } c - k \text{ хосын тоо}}{m_u - 1} \quad (2.15)$$

Нэгж тус бүр гарцаагүй $m_u(m_u - 1)$ ширхэг тохиролцлыг агуулна. (2.12), (2.13), (2.14) томъёонд байх зөрүүгийн функцийг $metric\delta_{ck}^2$ 4-р нөхцөлд ашиглана. Гэхдээ өгөгдлийн төрлөөс хамаарч хэмжүүр нь ($metric = [“nominal”, “ordinal”, “circular”, “interval”, “bipolar”, “ratio”]$) өөр өөр байна. Хэрэв синсетийн үгсийг Зөв, Буруу гэх мэт нэрийдсэн ангилалд оруулж санал өгвөл $nominal\delta_{ck}^2$ зөрүүгийн функц ашиглах шаардлагатай болно. Бусад зөрүүгийн функцүүдийг бодвол нэрийдсэн ангиллын зөрүүгийн функцэд параметр авдаггүй бөгөөд бусдаасаа хамгийн энгийн, хялбар илэрхийлэлтэй функц юм.

Хүснэгт 2.6 Нэрийдсэн ангиллын зөрүүгийн функц

Нэрийдсэн ангилал, нэрс:

$$nominal\delta_{ck}^2 = \begin{cases} 0 & \text{iff } c = k \\ 1 & \text{iff } c \neq k \end{cases}$$

	a	b	c	d	e
a	0	1	1	1	1
b	1	0	1	1	1
c	1	1	0	1	1
d	1	1	1	0	1
e	1	1	1	1	0

Бүлгийн дүгнэлт

Олны хүч бол компьютерийн гүйцэтгэж чадахгүй байгаа, нөгөө талаас хүний оюун зайлшгүй шаардлагатай ажлуудыг хүний оюуны тусламжтайгаар богино хугацаанд бага зардлаар гүйцэтгэх боломжийг олгож буй шинэ судалгааны арга бас хэрэгсэл болж байна. Энэ аргаар маш олон тооны хүмүүсийн оруулсан хувь нэмрийг нэгтгэж үнэлснээр үнэн мэдлэг гаргаж авч болохыг, ялангуяа орчуулгын ажилд мэргэжлийн түвшний орчуулгыг хийх боломжтойг зарим судлаачид нотлон харуулжээ. Тэд энгийн орчуулгын чанарыг үнэлэх маш олон аргуудыг боловсруулж хэрэгжүүлсэн бол онтологийг нутагшуулах ажилд үгийн сангийн орчуулгыг олны хүчээр гүйцэтгэх нь харьцангуй шинэ судалгааны ажил байна. Гэвч ялгамжит онтологийг олны хүчээр нутагшуулах аргачлал хөндөгдөөгүй хэвээр байна.

Олны хүчээр гүйцэтгэсэн ажлын үр дүнг жишиг сантай харьцуулж үнэлдэг. Гэвч эдгээр үнэлгээний аргуудад статистик магадлалын хэмжүүрийг ашиглан олны хүчээр гүйцэтгэсэн ажлыг хэр зөв болсныг үнэлэх, улмаар зөв ажлуудаас зөв үр дүнг гаргаж авах аргуудыг боловсруулаагүй байна. Үүнд Флайс каппа, Криппендорффийн альфа зэрэг олны дотоод санал нийцлийг олох статистик хэмжүүрийг ашиглан синсетийн орчуулгын чанарыг үнэлэх, чанартай орчуулсан синсетээс оновчтой үгийг сонгож үр дүнг нэгтгэх боломжтой гэж үзэж байна.

БҮЛЭГ 3. ОЛНЫ ХҮЧЭЭР ОНТОЛОГИ НУТАГШУУЛАХ АРГАЧЛАЛ

Энэ бүлэгт онтологи нутагшуулах ажлын даалгавар болон түүний шаардлага, онтологийг энгийн вэб хэрэглэгч болон мэргэжилтнүүдээр нутагшуулах хосолмол аргачлал болон түүний туршилт, үр дүнг танилцуулна. Энэ ажлаар бид ЕМС-ийн ойлголтын цөм болон хэлний цөмийг нутагшуулах аргачлалыг боловсруулсан бөгөөд үүнийг энэ бүлэгт онтологи нутагшуулах гэж ойлгоно.

3.1 Онтологи нутагшуулах ажил

Энэ бүлэгт онтологи нутагшуулах ажлын даалгавар, түүнд тохиолдох ялгамжийн төрлүүд, олны хүчээр хийх даалгаврын шаардлага, оролцогчдын хувь нэмрийг нэгтгэх олны санал нийцлийн арга, онтологийн чанарын үнэлгээний аргыг танилцуулна.

3.1.1 Ажлын даалгавар

ЕМС дахь онтологи нутагшуулах ажлаар [25]–[27] нутагшуулах элементүүдийг тодорхойлж шинээр дүйцэлгүй ойлголтыг ЕМС-д нэмж оруулсан болно. Эдгээр элементүүд ойлголтын цөм болон хэлний цөмийг хамардаг бөгөөд жагсаалтыг Хүснэгт 3.1-д үзүүлэв.

Хүснэгт 3.1 Нутагшуулах онтологийн элементүүд

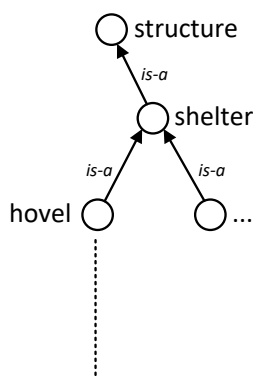
№	Элемент	Тайлбар
1.	Ойлголт	Класс, шинжийн нэр, холбоосны нэр, шинжийн утгыг илэрхийлэх формаль ухагдахуун
2.	Утгазүйн холбоос	Хоёр ойлголт хоорондын холбоо
3.	Утгалбар	Үгийн утгыг илэрхийлэх бөгөөд нэг үг болон энэ үгийг агуулж буй синсет хоорондын хэлхээг тодорхойлдог
4.	Үгийн аймгийн тэмдэглэгээ	Синсенийн үгийн аймгийн тэмдэглэгээ (нэр үг, үйл үг, тэмдэг нэр, дайвар үг г.м.)
5.	Үгзүйн холбоос	Хоёр утгалбар хоорондын холбоо
6.	Утгалбарын товч тайлбар	Синсенийн тайлбарыг товч илэрхийлэх үгийн сангийн нэгж
7.	Үглэвэр	Тодорхой нэг үгийн санд хамаарах үгийг илэрхийлнэ
8.	Үглэврийн хувилбар	Үгийн өөр хэлбэр
9.	Синсет	Тодорхой нэг үгийн санд хамаарах ойролцоо утгатай бүлэг үгс

10. Синсетийн тайлбар	Синсетийн утгыг илэрхийлэх бичвэр
11. Үг-утгазүйн холбоос	Хоёр синсет хоорондын холбоо
12. Утгалбарын эрэмбэ	Нэг үгийн утгалбаруудын эрэмбэ буюу дараалал
13. Синсетийн үгийн эрэмбэ	Синсетийн үгсийн эрэмбэ буюу дараалал
14. Жишээ өгүүлбэр	Синсетийн утгыг харуулах жишээ өгүүлбэр
15. Дүйцэлгүй ойлголт	Аливаа хэлний үгийн сангийн нэгжээр илэрхийлж чадахгүй ойлголт

Дээрх хүснэгтэд үзүүлсэн 15 элементүүдийн хоорондын холбоо хамаарлыг Зураг 1.6-д үзүүлсэн болно. Ойлголтоос бусад бүх элементүүд ойлголтоос эх авч холбогдоно. Жишээлбэл, ойлголт ямар нэг синсетээр илэрхийлэгдэнэ, тэр синсет олон үгсээс тогтоно, мөн тайлбартай байна, тэр синсетийн нэг үг олон утгалбартай байна, түүний нэг утгалбар өөр утгалбартай үгзүйн нэг холбоосоор холбогдож болно гэх мэт ойлголтоос эхтэй холбоосоор бүх элементүүд холбогдоно.

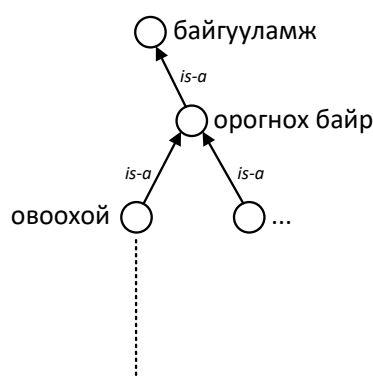
Онтологи нутагшуулах ажлын ерөнхий даалгавар бол эх А) онтологийг зорилгын Б) онтологид нийцүүлэн буулгах юм (Зураг 3.1). Үүнийг хэрэгжүүлэхэд хамгийн эхний ажил бол эх хэлний **синсетийг** зорилтот хэлний синсет рүү орчуулах юм.

А) эх онтологи



(noun) **hovel, hut, hutch, shack, shanty**
-- small crude shelter used as a dwelling

Б) зорилгын онтологи



(нэр) **овоохой, урц, оромж** – орон байр
шиг ашигладаг жижиг болхи орогнох байр

Зураг 3.1 Онтологи нутагшуулах ажлын ерөнхий даалгавар

Дээрх зурагт *hovel* гэх **ойлголтыг** илэрхийлэх эх хэлний синсет {hovel, hut, hutch, shack, shanty}-ийг зорилтот хэлний синсет {овоохой, урц, оромж} гэж орчуулсан. Ингэхдээ **синсетийн үгсийн эрэмбийг** зааж өгөх шаардлагатай. Энд овоохой – 1, урц – 2, оромж – 3 гэж өсөх дарааллаар эрэмбэлэгдэнэ. Мөн тухайн синсетийн **үгийн**

аймгийн тэмдэглэгээг онооно. Ихэвчлэн эх хэлний үгийн аймгийн тэмдэглэгээтэй ижил байдаг. Учир нь нэрийдсэн ойлголтыг илэрхийлэх синсет эх хэлэндээ нэр үг тэмдэглэгээтэй байдаг бол энэ нь мөн зорилтот хэлэнд нэр үг тэмдэглэгээтэй байдаг. Үйлийг илэрхийлж байвал үйл үг тэмдэглэгээтэй байдаг. Дараа нь **синсетийн тайлбарыг** орчуулна. Энэ нь зорилтот хэлэнд тухайн ойлголтыг илэрхийлж чадах, хүн уншихад ойлголт нь дүрслэгдэх тайлбар бичвэр юм. Энэ нь бүтэн өгүүлбэр бус өгүүлбэрийн дэд хэсэг байдаг. Эх онтологид *hovel* бол *shelter* гэсэн хоёр ойлголтын хоорондох **утгазүйн холбоос** нь *овоохой* бол *орогнох* байр гэж зорилгын онтологид ижил бууна. Харин утгалбарын хувьд овоохой гэдэг үг Монгол хэлэнд 2 утгатай бөгөөд хоёрдугаар утга нь *муу гэр* юм. Энэ тохиолдолд хамгийн их хэрэглэгддэг *жижиг болхи орогнох байр* гэдэг утга нь **утгалбарын эрэмбийн** хувьд 1, *муу гэр* гэдэг утга нь 2 болно.

Хэрэв тухайн ойлголтыг зорилтот хэлэнд илэрхийлж чадахгүй бол энэ нь **дүйцэлгүй ойлголт** болох бөгөөд шууд орчуулагдах боломжгүй. Дүйцэлгүй ойлголт ялгамжийн гол төлөөлөгч боловч аливаа ялгаатай байдлыг ойлголтын, синсетийн болон утгалбарын түвшинд тодорхойлж болно.

3.1.2 Ялгамжийн төрлүүд

Онтологи нутагшуулах нь энгийн ажил биш, хэл шинжээчид, ялангуяа танихуйн хэл шинжээчид болон онтологийн инженерүүдийн хамтын ажил юм. Нутагшуулах ажил бодит ертөнц дээрх нэгж-объектууд болон ойлголтуудыг тухайн орон нутагт хэрхэн төсөөлж ойлгодог байдалд түшиглэн явах хэрэгтэй. Өөрөөр хэлбэл, орчлон дээрх нэг ойлголтыг хоёр өөр соёлд хоёр өөр өнцгөөр хүлээн авч тайлбарладаг байж болно. Ийм учраас энэ ажил дан ганц шууд (үгчилсэн) орчуулгаар хийгдэхгүй нь тодорхой юм. Бид ялгамжийг 3 түвшинд тодорхойлсон [25].

Ойлголт

ЕМС-д ойлголтыг орчлон дээрх нэг зүйл гэж авч үзсэн. Түүнийг олон хэл дээр янз бүрээр илэрхийлдэг. Нэг хэлний хувьд ч гэсэн нэг ойлголтыг олон нэр томъёогоор (ойролцоо утгатай үгс) илэрхийлдэг. Олон ойлголтыг нэг нэр томъёогоор (сацрал утгатай үгс) илэрхийлдэг байна. Жишээ нь, *valley*, *dale*, *hollow* гэсэн 3 ойлголтыг монгол хэлэнд нэг үгээр илэрхийлдэг.

valley -- a long depression in the surface of the land that usually contains a river

dale -- an open river valley (in a hilly area)

hollow -- a small valley between mountains

ЕМС-д *dale* болон *hollow* ойлголтууд *valley* ойлголтын доор байрлах хүү (child) ойлголтууд юм. Энэ тохиолдолд эдгээр ойлголтуудыг орчуулахад зорилтот хэл дээр сацрал утгатай үгсийг ихэсгэх болно. Сацрал утгатай үгс гэдэг нь нэг үгийн цөм утгаас салбарлаж олон утга илэрхийлэх бөгөөд утгууд нь хоорондоо холбоотой байх үгс юм.



Зураг 3.2 Сацрал үгсээр илэрхийлэх ойлголт

Бид ийм орчуулгыг орчуулах хэрэгтэй учир нь Монгол соёл хүмүүс эдгээр ойлголтуудаар илэрхийлэх нэгж-объектуудыг ялган ойлгож чадах юм.

Дүйцэлгүй үг бол нэг төрлийн ялгамж юм. Жишээ нь, энэ диссертацийн 1.4.2 хэсэгт заасан *parish* үгийн өөр утга *the local subdivision of a diocese committed to one pastor* Монгол хэлэнд бас дүйцэлгүй үг болно. Эх (Э) хэлнээс зорилгын (З) хэлэнд дүйцэлгүй ойлголт болох хувилбарыг Зураг 3.3 (а)-д үзүүлэв.



а) зорилгын хэлэнд дүйцэлгүй үг

б) эх хэлэнд дүйцэлгүй үг

Зураг 3.3 Дүйцэлгүй ойлголтын хувилбар

Дүйцэлгүй үг бол хэлний нэг онцлог тул бүх хэлэнд тохиолдоно. Зорилтот хэлнээс эх хэлэнд бас дүйцэлгүй үг байж болно. Жишээ нь, Монгол хэлний *бууц* болон *буурь* гэдэг үгс нь англи хэлэнд дүйцэлгүй үг болно. *бууц* гэдэг үгийн утгыг англи хэлэнд *an area of dried and accumulated manure where a nomadic family was living* гэж тайлбарлаж болно. *буурь* гэдэг үгийг *a round shaped spot where a nomadic yurt was built* гэж тайлбарлаж болно. Эдгээр үгсийг англи хэлний үгийн сангийн ямар нэг нэгжээр илэрхийлж чадахгүй байгаа тул үүнийг дүйцэлгүй үг гэж ойлгоно. Эдгээр нь Монголын нүүдэлчин амьдралын хэв маягаас үүдэлтэй ойлголтууд бөгөөд англи

хэлээр ярьдаг бусад соёлд хэрэглэдэггүй ойлголт юм. Энэ хувилбарыг Зураг 3.3 (б)-д дүрслэв.

Дүйцэлгүй үг нь заримдаа синсетийн тайлбар дотор бас тохиолддог. Жишээ нь, *piers* гэдэг нэр томъёо нь *Romanesque architecture* (Романы архитектур) ойлголтын тайлбарт орсон байна. Энд *pier* гэдэг нь нэг төрлийн багана (column) шиг боловч баганаас ялгаатай ойлголт бөгөөд зорилтот хэлэнд ийм ойлголтыг хэрэглэдэггүй бол энэ дүйцэлгүй үг болно. Энэ тохиолдолд синсетийн орчуулгад тухайн үгийг чөлөөт бичвэрээр илэрхийлнэ.

Romanesque architecture -- ... *characterized by round arches and vaults and by the substitution of **piers** for columns and profuse ornament and arcades*

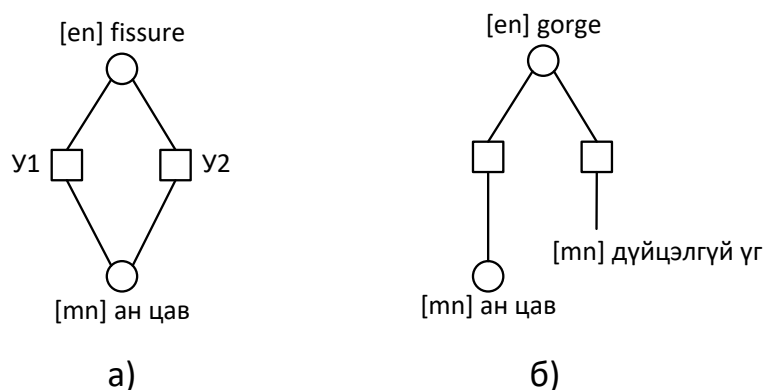
Утгалбар

Зарим үгийн олон утга хоорондоо бага зэргийн ялгаатай байдаг. Жишээ нь, англи хэлний *fissure* гэдэг үг дараах хоёр утгатай байна.

[Y1] crack, cleft, crevice, **fissure**, scissure -- *a long narrow opening*

[Y2] **fissure** -- *a crack associated with volcanism*

Дээрх хоёр ойлголт эх газрын нам дор газар гэдэг ойлголттой төрөл (hyponym) ойлголт бөгөөд энэ хоёр зорилтот хэлэнд бас ижил нэрээр (homonym) илэрхийлэгдэж болно. Энэ үзэгдлийг Зураг 3.4 (а)-д үзүүлэв. У – утгалбар болно.



Зураг 3.4 Үгийн утгалбарын ялгамж

Заримдаа сацрал утгатай үгийн зарим утгалбарынхаа хувьд дүйцэлгүй үг байж болно. Жишээ нь, *gorge* гэдэг үг хоёр утгалбартай бөгөөд түүний нэг нь зорилтот хэлэнд дүйцэлгүй үг болно (Зураг 3.4 (б)). Зураг 3.4-т en – English, mn – Mongolian гэж олон улсын ISO 639-1 хэлний кодын стандартын товчлолыг ашиглав.

Синсет

Эх хэлний синсетийн үгс зорилтот хэлэнд шууд орчуулагдаж болохыг өмнөх дэд бүлэгт өгүүлсэн. Харин тэдний зарим нь орчуулагдахгүй байж болно. Жишээ нь, англи хэлний синсет *peak (the top point of a mountain or hill)* 6 үгтэй. Тэдний 3-ыг Монгол хэлэнд тухайн ойлголтыг илэрхийлэх утгаар орчуулах боломжгүй байна.

1 peak → оргил

2 crown

3 crest

4 top → орой

5 tip

6 summit → ноён оргил

Өөрөөр хэлбэл, 6 үгтэй синсетийг орчуулах явцад 3 үгийг англи үгийн орчуулгаас олж чадсан бөгөөд үлдсэн 3 үг *crown, crest, tip* зэрэг үгийн орчуулгаас энэ синсетэд нэмж оруулах боломжтой үгс байхгүй юм. Гэхдээ нь *peak, top, summit* зэрэг 3 үгс Монгол хэлэнд *оргил, орой, ноён оргил* зэрэг үгстэй харгалзана гэсэн үг биш юм.

Синсетийн тайлбарын зарим хэсгийг үгийн шууд орчуулгын оронд маш ойр утгатай эсвэл ижилхэн утгатай үгсийг ашиглан орчуулж болно. Хэдийгээр синсетийн тайлбарын шууд орчуулгыг хийх нь оновчтой боловч зарим тохиолдолд шууд орчуулга нь боломжгүй байна. Жишээ нь, доорх синсетийн тайлбарт “near a shore” гэдэг хэсгийг Монгол хэлний орчуулгаас хасаж орчуулгыг өөрчилсхийж (paraphrasing) хийсэн болно. Учир нь англи хэлний *bank* болон *shore* гэдэг ойлголтуудыг Монгол хэлэнд ялгаж ярьдаггүй.

[en] oceanic sandbank -- a submerged bank of sand **near a shore**, can be exposed at low tide

[mn] далайн элсэн эрэг -- *шунгаж орсон далайн элсэн эрэг, далайн давалгааны намхан хаялганд үзэгддэг* (gl. a submerged sea bank of sand, visible at low tide)

Мөн синсетийн жишээ өгүүлбэрийн хувьд ч гэсэн **өөрчилсхийж орчуулах** эсвэл шинэ өгүүлбэр оруулах эсвэл өгүүлбэрийг сольж болно. Учир нь жишээ өгүүлбэр бол тухайн ойлголтыг илэрхийлэхүйц, синсетийн аль нэг үг оролцсон байх нь оновчтой. Жишээ нь, олны танил нэр жишээ өгүүлбэрт оролцож тухайн ойлголтын талаар

тодорхой төсөөлөл өгч байгаа бол тухайн орон нутагт ойр олны танил нэрээр солих нь илүү сайн тайлбарыг бий болгоно. Алфийн нурууны ноён оргил бол Монт Бланк гэсэн өгүүлбэрийг Монгол Алтайн нурууны ноён оргил бол Хүйтэн оргил юм гэж өөрчилж болно. Цаашилбал, синсетийн тайлбарт хэмжих нэгжийн тэмдэглэгээ, жишээ нь, рН орвол тухайн орон нутагт хэрэглэдэг эсвэл олон улсын тэмдэглэгээг ашиглах нь зүйтэй. Жишээ нь, огноо болон цагийн формат, хэмжих нэгж болон валютын тэмдэглэгээ зэргийг тухайн орон нутгийн хэмжих нэгж рүү шилжүүлбэл илүү тохиромжтой. 5 инч Монголын ашигладаг MKS (Meter Kilogram Second) системийн хувьд 12.7 сантиметр болно. Ийм үгс зөвхөн синсетийн тайлбарт тохиолддог. Гэвч ийм үгс заримдаа таатай биш байдаг. Учир нь бутархай тоог хүмүүс тогтооход төвөгтэй юм.

3.1.3 Олны хүчээр хийх ажил

Онтологийг нутагшуулахад олны хүчээр хийлгэх ажлууд синсет орчуулах, синсетийн үгийн эрэмбийг тодорхойлох, дүйцэлгүй ойлголтыг тэмдэглэх, үгийн утгалбарын эрэмбийг тодорхойлох, шинэ ойлголт буюу эх хэлэнд дүйцэлгүй ойлголтыг олох, түүний утгазүйн холбоосыг тогтоох гэх мэт олон ажлууд бий. Гэвч эдгээр ажлуудаас аль ажлуудыг ямар үнэлэмжтэй оролцогчдоор гүйцэтгүүлэхийг тодорхойлох ёстой. Тухайлбал, синсетийн үгс болон тайлбарыг орчуулах ажлыг бага үнэлэмжтэй оролцогчид буюу энгийн вэб хэрэглэгчдээр гүйцэтгүүлэх боломжтой гэж үзэж байна. Учир нь мэргэжлийн түвшний орчуулгыг мэргэжлийн бус хүмүүсээр гүйцэтгүүлэх боломжтойг диссертацийн 2.2-р бүлэгт дурдсан ажлууд баталж байгаа билээ. Гэтэл үгийн утгалбарын эрэмбийг тодорхойлох нь өмнө судлагдаагүйгээс гадна айгаас хамаарч үгс өөр өөр эрэмбэ байж болдог тул ямар аргачлалаар тодорхойлох судалгааны бас нэг асуудал юм. Иймд утгалбарын эрэмбийг тодорхойлох нь хэцүү ажил билээ. Түүнчлэн шинэ ойлголт нэмэх нь зорилгын онтологиос эх онтологид дүйцэлгүй ойлголтыг олж илрүүлэх ажил юм. Мөн түүний холбоосыг тодорхойлох нь өндөр үнэлэмжтэй оролцогчдын мэдлэг шаардсан ажил болох нь гарцаагүй.

Хүснэгт 3.2 Олны хүчээр хийх ажлууд

№	Элемент	Хийх ажил	Ажилчид
А. Үгийн утгалбар			
1	Синсетийн үгс	орчуулах	бага
2	Зорилтот хэл дээрх синсетийн үг	нэмэх	бага

3	Синсетийн тайлбар	орчуулах	бага
4	Синсетийн жишээ өгүүлбэр	орчуулах, нэмэх	бага
5	Үглэврийн хувилбар	тодорхойлох	бага
6	Үгийн утгалбарын эрэмбэ	тогтоох	өндөр
7	Синсетийн үгийн эрэмбэ	тогтоох	бага
8	Утгалбарын товч тайлбар	орчуулах	бага
В. Холбоос			
9	Утгалбар / Синсет / Ойлголтын холбоос	тодорхойлох	өндөр
10	Холбоосны төрөл	орчуулах	өндөр
11	Хэлнээс хамаарсан холбоосны шинэ төрөл	тодорхойлох, орчуулах	өндөр
С. Дүйцэлгүй ойлголт			
12	Зорилтот хэлний дүйцэлгүй ойлголт	тодорхойлох	бага
13	Зорилтот хэлний дүйцэлгүй ойлголтын тайлбар	орчуулах	бага
14	Эх хэлэнд дүйцэлгүй ойлголт	тодорхойлох	өндөр
15	Эх хэлний дүйцэлгүй ойлголтын тайлбар	орчуулах	өндөр
16	Эх хэлэнд дүйцэлгүй ойлголтыг илэрхийлэх шинэ үг	тодорхойлох	өндөр

Иймд Хүснэгт 3.2-т онтологи нутагшуулалтын ямар ажлыг ямар үнэлэмжтэй оролцогчид гүйцэтгэж чадахыг тодорхойлов. Энд *орчуулах* гэдэг нь эх хэл дээр байгаа үгс, тайлбар зэрэг элементүүдийг зорилтот хэлэнд дүйцүүлэн буулгах; *нэмэх* нь өмнө байсан зүйл дээр шинээр нэмж тодорхойлох; *тогтоох* гэдэг нь бий болсон зүйлийг өөрчлөн тодорхойлох; *тодорхойлох* нь шинээр олж мэдсэний үндсэн дээр эсвэл тийм байх ёстойг мэдсэн тохиолдолд нэмэх үйлдэл юм. Жишээ нь, зорилтот хэлэнд дүйцэлгүй ойлголт мөн болохыг олж мэдсэний дараа үүнийг тэмдэглэх нь *тодорхойлох* үйлдэл болно. Эсвэл утгалбар хооронд байх ёстой холбоосыг олох нь үүний нэг жишээ. Синсетийн жишээ өгүүлбэрийг зохиож бичвэл энэ нь нэмэх үйлдэл болно. Мөн дээрх хүснэгтийн 2 дугаар мөрөнд байгаа Зорилтот хэл дээрх синсетийн үг гэдэг нь зорилтот хэл дээр шинээр нэмж оруулах үгийг хэлнэ. 5-р мөрөнд буй Үглэврийн хувилбар гэдэг тухайн үгийн дүрмийн бус аргаар хувирсан хэлбэрийг хэлнэ. Жишээ нь, англи хэлэнд wife – wives, wolf – wolves юм. 13, 15-р мөрөнд дүйцэлгүй ойлголтын тайлбар гэдэг нь синсетийн тайлбараас ялгаатай юм. Учир нь зорилтот хэл дээрх дүйцэлгүй ойлголт эх хэлний ямар нэг синсеттэй холбоотой бөгөөд

угтаа өөрөө бол синсет биш тул үүнийг ойлголтын тайлбар гэж синсетийн тайлбараас ялгаж байгаа хэрэг юм.

3.2 Мэргэжилтнүүдээр онтологи нутагшуулах аргачлал

Мэргэжлийн ажлын онлайн зах UpWork¹⁸, Freelancer¹⁹ зэрэг вэб сайтуудад мэргэжлийн ажил хийх, мэргэжлийн ажил хийлгэх боломжтой. Энд програм хангамж болон вэб хөгжүүлэх, медиа зөвлөгөө өгөх, зохиол бичих гээд маш олон төрлийн ажлууд байдаг. Гэвч эдгээр нь мэргэжилтэн хөлсөлж ажиллуулаад төлбөр төлөх маягаар ажилладаг. Олон шат дараалалтай, төрөл бүрийн мэргэжилтэн шаардсан ийм ажлыг олны хүчний ажиллах зарчимд шилжүүлж богино хугацаанд чанартай үр дүн гарган авах тухай судалгаанууд байдаг. Тухайлбал, маш богино хугацаанд гялс баг (flash) бүрдүүлэн гар утасны програмын хэрэглэгчийн интерфэйсийн загвар, богино хэмжээний хүүхэлдэйн кино зэргийг олны хүчээр гүйцэтгүүлсэн [68]. Үүний нэгэн адил нэгэнт онтологи нутагшуулах нь айн мэргэжилтэн, онтологи инженерийн мэдлэг шаарддаг тул энгийн бага үнэлэмжтэй оролцогчдоор хийлгэж болмооргүй ажлыг мэргэжлийн хүчээр (expert sourcing) гүйцэтгүүлэх боломжтой. Энэ бүлэгт мэргэжлийн хүчийг ашиглан онтологийг нутагшуулах аргыг тайлбарлана.

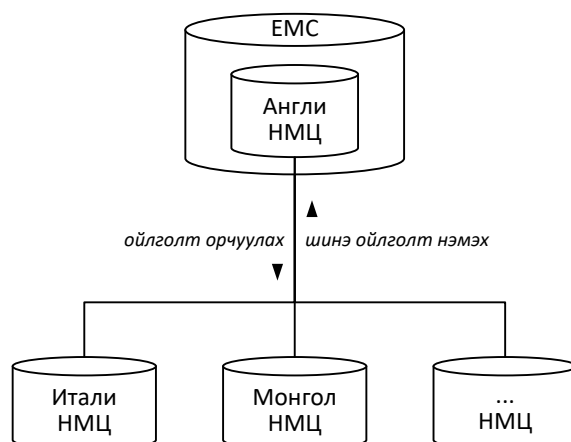
3.2.1 Нутагшуулах үйл явц

Онтологийг нутагшуулах ажлыг онтологийн ойлголтуудыг илэрхийлэх үгийн санг орчуулах ажил болгон хувиргасан тухай өмнөх дэд бүлэгт танилцуулсан. Орчуулгын ерөнхий санаа нь эх хэл дээр байгаа ямар нэг элементийг авч зорилтот хэлэнд түүнд тохирох оновчтой элементийг үүсгэж өгөх юм. Гэвч нэг ойлголтыг нэг синсет илэрхийлдэг бөгөөд бусад элементүүд үүнтэй уялдаатай үүсэх учир гол зангилаа элемент нь синсет билээ. Иймд синсетийг орчуулах нь хамгийн чухал даалгавар юм. Бид англи хэлийг онтологи нутагшуулах ажил бүрд эх хэлээр сонгосон. Учир нь энэ хэл дэлхийд хамгийн өргөн хэрэглэгддэг 2 дахь хэл бөгөөд шинжлэх ухаан, эрдэм шинжилгээний салбарт нийтлэг ашигладаг хэл юм. Нутагшуулалтын үйл явцыг хэрэгжүүлэхэд EMC-ийн англи хэл дээрх хэлний цөмийг (NLCore) Нутгийн мэдлэгийн цөм (НМЦ - LKC - Local Knowledge Core) гэгдэх санд хуулж бэлдэнэ. Нутгийн мэдлэгийн цөм зорилтот хэлний онтологийг хадгалдаг сан юм (Зураг 3.5).

¹⁸ <https://www.upwork.com>

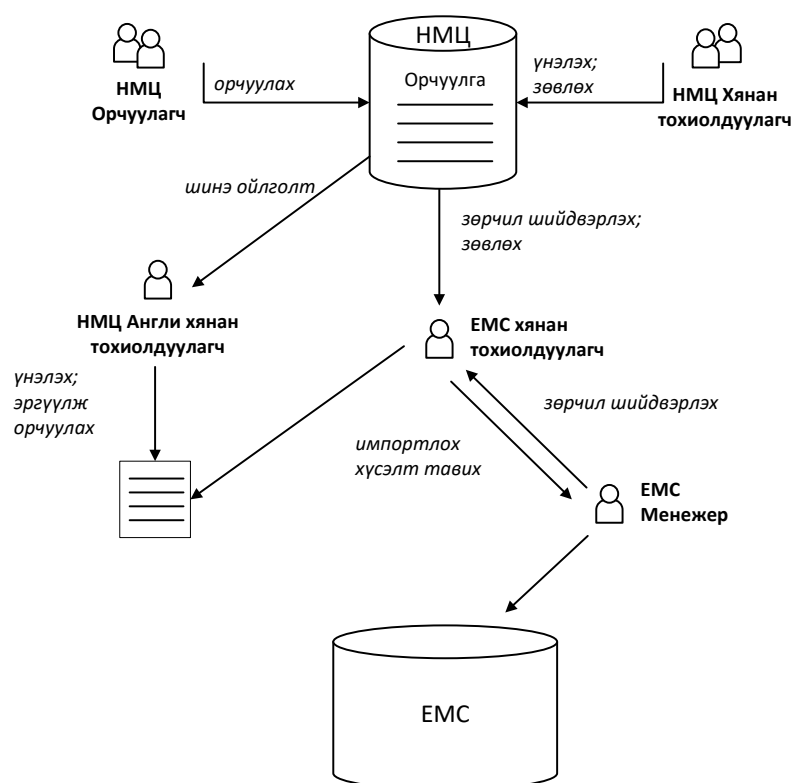
¹⁹ <https://www.freelancer.com>

Энд эх хэлнээс ойлголтыг орчуулах (concept translation), Нутгийн мэдлэгийн цөмөөс шинэ ойлголт EMC-д нэмэх (concept generation) гэсэн үндсэн хоёр үйлдэл хийгдэнэ.



Зураг 3.5 Нутгийн мэдлэгийн цөм ба Ертөнцийн мэдлэгийн сангийн уялдаа

Нутгийн мэдлэгийн цөмд синсетийг орчуулах аргачлалыг макро алхмаар харуулав (Зураг 3.6).



Зураг 3.6 Нутгийн мэдлэгийн цөмд синсетийг орчуулах макро алхамууд

- 1(а) **НМЦ орчуулагч** (LKC Translator) эх хэл дээр байгаа синсетийг сонгон авч түүний утгын талаар маш сайн ойлголт авна. Ингэхдээ бусад эх сурвалжууд,

тухайлбал, вэб дэх зураг, видео, тайлбар толь бичгийн тайлбар зэргийг ашиглаж болно.

- 1(б) **НМЦ орчуулагч** зорилтот хэл дээр синсенийн үг(с)ийн хамгийн оновчтой орчуулгыг тодорхойлно. Энд синсенийн үг(с) гэдэгт үгийн сангийн нэгж буюу үг, нийлэмж, хэлц үг байна. Үүнд чөлөөт бичвэрийг хүлээж авахгүй. Хэрэв синсенийн утгыг илэрхийлэх ямар нэг үг тодорхойлж чадахгүй бол орчуулагч дүйцэлгүй үг гэж тэмдэглэнэ. Хэдий тийм байсан ч орчуулагч үргэлж синсенийн тайлбарыг орчуулах ёстой.
- 1(в) **НМЦ хянан тохиолдуулагч (LKC Validator)** синсенийн үг(с) болон тайлбарыг үнэлнэ. НМЦ орчуулагчийн орчуулгатай санал нийлэхгүй байвал яагаад санал нийлэхгүй байгааг тайлбарлан санал шүүмж өгнө. Хэрэв дүйцэлгүй үг байвал хянан тохиолдуулагч дүйцэлгүй үг болохыг нь баталгаажуулах эсвэл тухайн синсетэд тохирох үг(с)ийн орчуулгыг тодорхойлно.
- 1(г) **НМЦ орчуулагч** синсенийн орчуулгад ирсэн санал шүүмжийг хүлээн авсны дараа шаардлагатай бол орчуулгыг шинэчилж засварлана. Хэрэв НМЦ хянан тохиолдуулагчийн санал шүүмжтэй санал нийлэхгүй байвал учир шалтгааныг тодорхой тайлбарлан хариу санал шүүмж өгнө.
- 1(д) **НМЦ хянан тохиолдуулагч** шинэчилсэн орчуулгыг дахин үнэлнэ. Хэрэв санал нийлэхгүй бол дахин санал шүүмжийг хэлний орчуулагчид өгнө (Алхам 1(г)). Ийм байдал хэд хэдэн удаа давтагдсан ч санал нийлэхгүй бол хоёр дахь НМЦ хянан тохиолдуулагч орчуулгыг үнэлэх ажлыг авч гүйцэтгэнэ. Хэрэв хянан тохиолдуулагч болон орчуулагч хоёр санал нийлсэн гэж үзвэл тухайн синсенийг үнэлэх ажил дуусгавар болно.
- 2 **НМЦ орчуулагч** орчуулах явцад шинэ ойлголт буюу дүйцэлгүй ойлголт байж болох эсвэл эх хэл дээр орхигдсон буюу эх хэлэнд орох ёстой ойлголт олж илрүүлбэл зорилтот хэл дээр шинэ синсет болон түүнд харгалзах эх хэлний синсенийг үүсгэнэ. Энэ тохиолдолд **НМЦ англи хянан тохиолдуулагч (LKC English Validator)** зорилтот хэлнээс гарч ирсэн шинэ ойлголтыг илэрхийлэх эх хэлний синсенийг үнэлнэ. Хэрэв энэ нь эх хэлэнд дүйцэлгүй ойлголтыг илэрхийлэх бол зорилтот хэлний синсенийн тайлбарыг англи хэл рүү орчуулна. Хэрэв энэ нь эх хэлэнд орхигдсон синсет байвал эх хэлэнд шинээр синсенийн орчуулгыг баталгаажуулна.

- 3 **ЕМС Хянан тохиолдуулагч (UKC Validator)** 1 болон 2-р алхмын үр дүнг хэлний болон ЕМС-ийн талаас үнэлнэ. Энэ хянан тохиолдуулагч аль болох цөөн удаагийн үйлдлээр НМЦ хянан тохиолдуулагч болон НМЦ англи хянан тохиолдуулагчтай (шаардлагатай бол) харилцаж аливаа алдааг залруулах, асуудлыг шийдвэрлэнэ. Эцэст нь ЕМС-д орчуулгыг импортолж оруулах хүсэлтийг **ЕМС Менежер**т тавина.
- 4 **ЕМС Менежер (UKC Manager)** ЕМС-д импортолж оруулахаас өмнө нутагшуулах элементтэй холбоотой алдааг автоматаар шалгаж үнэлэх програмыг ажиллуулна. Хэрэв алдаатай бол ЕМС Хянан тохиолдуулагчийн тусламжтай алдааг засна. Энэ менежер бас зорилтот хэлнээс гарч ирж буй шинэ ойлголтуудыг зөвшөөрөх эсвэл зөвшөөрөхгүй байх эрхтэй.

Газарзүйн байршлаас харвал ихэнх тохиолдолд НМЦ-ийг тухайн хэлээр ярьдаг улс оронд хөгжүүлж байхад ЕМС-г дундаа хөгжүүлнэ. Энэ менежментийг НМЦ Хянан тохиолдуулагч болон ЕМС Менежер нар голлон гүйцэтгэнэ. Ийм тархсан үйл ажиллагаа болон операторууд нутгийн ялгамжийг олж илрүүлэн бүрдүүлж байхад цор ганц ЕМС-д төвлөрсөн зохицуулалт хийж байх юм. Учир нь тухайн ойлголт аль хэл соёл байхаас үл хамаарч ЕМС-д цор ганц байна. Иймд ЕМС бол нэг дэлхийг илэрхийлж харин өөр өөр хэлээр тэрхүү дэлхийг өөр өөр өнцгөөс харах юм. Үүний үр дүнд ялгамжийг ердийн хүмүүний хэлнүүдээс ЕМС-ийн формаль ойлголтуудад буулгаж нэг ЕМС-ийг байгуулах юм.

Дараах дэд бүлгүүдэд ойлголтыг орчуулах, шинэ ойлголт нэмэх алхмуудыг тайлбарлаж жишээгээр үзүүлнэ.

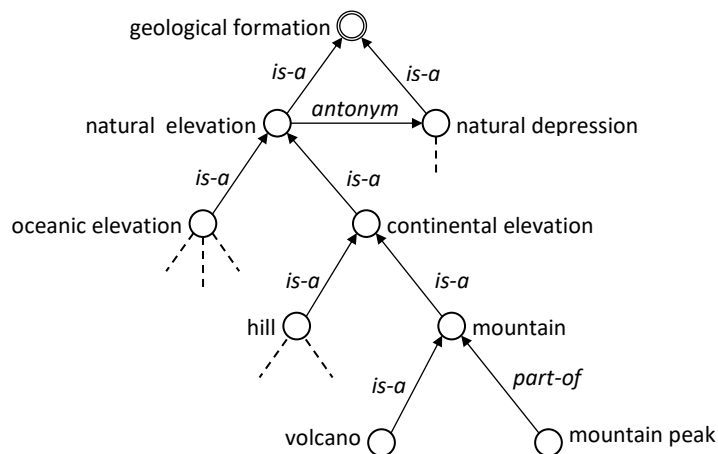
3.2.1.1 Ойлголтыг орчуулах

Ойлголтыг орчуулах ажлыг эхлүүлэхийн тулд НМЦ орчуулагчид онтологиос дэд модыг оноож өгнө. Жишээ нь Зураг 3.7-д орон зайн айн газарзүйн тогтоц (geological formation) фасетийн дэд хэсгийг үзүүлэв. Хэрэв энэ фасетаас *mountain* дэд модыг орчуулагчид оноовол орчуулагч дараах синсетүүдийг орчуулах даалгавартай болно гэсэн үг.

a) mountain, mount -- *a land mass that projects well above its surroundings; higher than a hill*

b) volcano -- *a mountain formed by volcanic material*

в) mountain peak -- *the summit of a mountain*



Зураг 3.7 Орон зайн айн газарзүйн тогтоц фасетийн дэд хэсэг

Энэ дэд модыг орчуулах энгийн зам бол 1, 3 болон 4 алхамуудыг хэрэгжүүлснээр тодорхойлогдоно. Дараах жишээгээр *mountain* синсетийг Монгол хэл рүү хэрхэн орчуулж буйг макро алхам бүрийн ард штрих (') тэмдэглэгээ тавьж харуулав.

1(а)' **НМЦ орчуулагч** *mountain* синсетийн тайлбарыг уншиж ойлголт авна. Энэ орчуулагч энэ ойлголт бол том хэмжээтэй, их эзлэхүүнтэй газар бөгөөд түүнийг хүрээлэн буй бусад газарзүйн тогтоцуудаасаа дээш өргөгдсөн байна гэдгийг ойлгоно. Мөн энэ нь толгодоос (*hill*) өндөр гэдгийг ч ойлгож чадна. Дараа нь орчуулагч яг ийм ойлголт эсвэл ухагдахууныг илэрхийлэх нэр томъёо Монгол хэлэнд байгаа эсэхийг шалгаж үзнэ. Ямар нэг хоёрдмол утга, эсвэл үүнтэй төстэй боловч бага зэрэг ялгаатай эргэлзүүлсэн ойлголтууд санаанд буувал бусад эх сурвалжуудаас (жишээ нь, вэб сайт, зураг, видео) нарийн ялгааг олж тогтоох шаардлагатай болно.

1(б)' **НМЦ орчуулагч** *mountain* синсетийг уул гэж Монгол хэлэнд орчуулна. Хэрэв энэ нь үгийн сангийн нэгжээр илэрхийлэгдэж байвал түүнийг дүйцэлгүй үг гэж тэмдэглэхгүй. Энэ тохиолдолд уул гэдэг үг Монгол хэлний үгийн сангийн нэгж мөн тул дүйцэлгүй үг болохгүй. Синсетийн тайлбарыг *эргэн тойрон буюу хүрээлэн буй орчноосоо дээш өргөгдөн гарсан өндөрлөг газар; толгодоос өндөр* (англи хэл рүү буцааж буулгавал: *a high land raised above and elevated from its surroundings and all-around; higher than hills*) гэж орчуулна.

1(в)' **НМЦ хянан тохиолдуулагч** энэ ойлголтыг илэрхийлэх уул гэдэг үгийг хүлээн зөвшөөрнө. Харин синсетийн тайлбарын орчуулгыг *эргэн тойрны орчноосоо дээш өргөгдсөн өндөрлөг газар; дов толгодоос өндөр*; (*a high land raised above*

from its surroundings; higher than hills) гэж сайжруулах санал өгөв. Энд хянан тохиолдуулагч *хүрээлэн буй, гарсан* зэрэг үгсийг тайлбараас хассан байна. Учир нь эдгээр үгсгүйгээр энэ синсенийн тайлбар унаган хэлтнүүдэд ойлгомжтой байж чадна.

1(г)' **НМЦ орчуулагч** *mountain* ойлголтын орчуулгад ирсэн санал шүүмжийг хүлээж авна. НМЦ Хянан тохиолдуулагчийн өгсөн саналын дагуу орчуулгыг шинэчилнэ.

1(д)' Нэгэнт **НМЦ орчуулагч** саналыг хүлээн зөвшөөрсөн тул энд ямар нэг санал зөрөлдөөнтэй зүйлс байхгүй болно. Иймд хянан тохиолдуулагч дараагийн алхам руу орчуулгыг явуулна.

3' **ЕМС Хянан тохиолдуулагч** синсенийн үгс болон тайлбарыг үнэлнэ. Энд ЕМС болон НМЦ-ийн талаас ямар нэг харшилдах зүйлгүй тул ЕМС Менежерт энэ ойлголтыг ЕМС-д импортлохыг асууна.

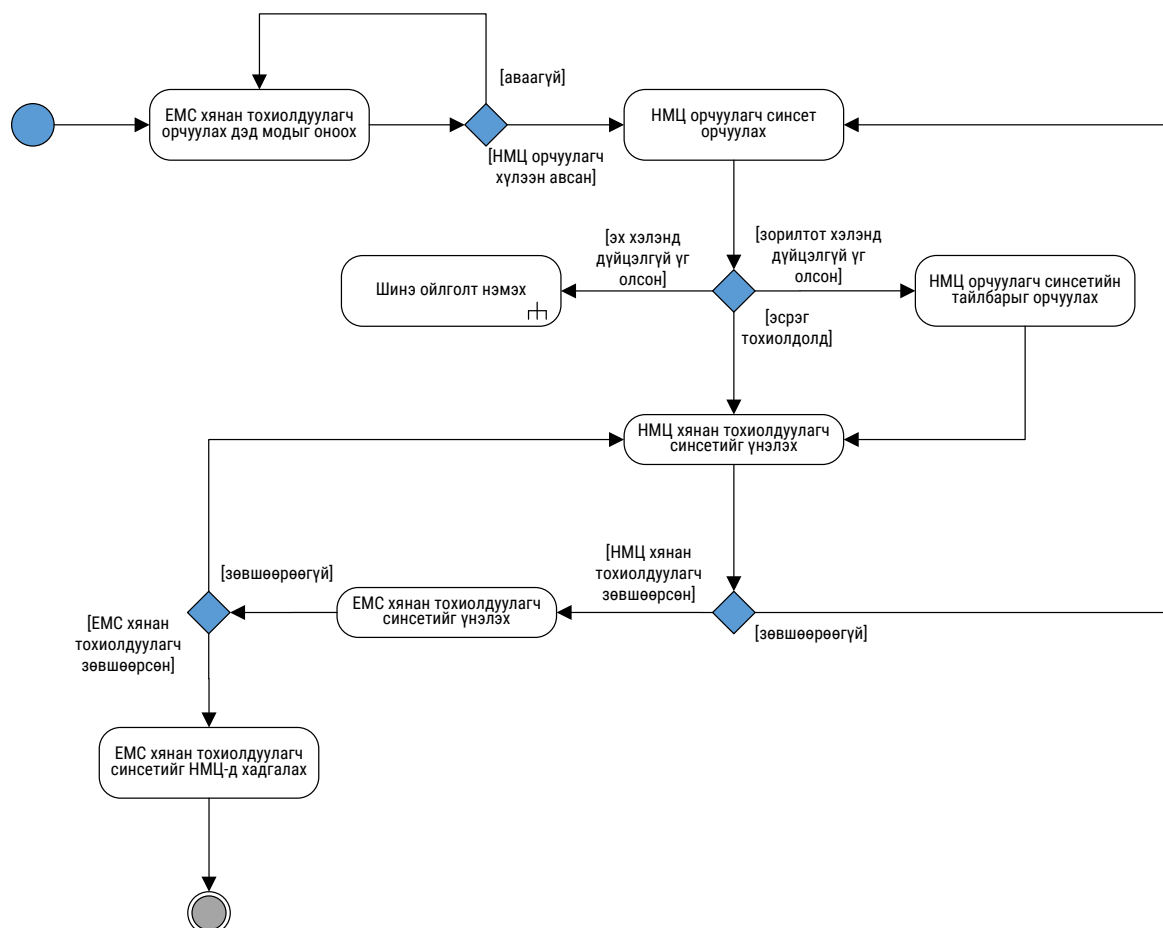
4' Төгсгөлд нь, **ЕМС Менежер** автомат алдаа шалгагчаар үүнийг оруулахад ямар нэг алдаа гараагүй тул ЕМС-д энэ орчуулгыг импортолж оруулна.

Үүнтэй төстэйгөөр *volcano, mountain peak* зэрэг ойлголтуудыг орчуулан ЕМС-д нэгтгэж болно.

Энэ үйл явцыг ерөнхий зураглалыг Зураг 3.8-д үзүүлэв. Энд НМЦ орчуулагч 3 боломжит үйлдлийг хийх боломжтой.

1. Тухайн ойлголт зорилтот хэлэнд дүйцэлгүй үг гэж тодорхойлох (lexical gap in target language)
2. Эх хэлэнд дүйцэлгүй үг олох (tentative lexical gap in source language)
3. Энгийн орчуулга гүйцэтгэх (else)

2 дахь үйлдлийн хувьд энэ нь шинээр ойлголт үүсгэх үйл явц болох ба энэ нь тусдаа процессын дагуу явагдана.



Зураг 3.8 Ойлголтыг орчуулах үйл явцын диаграм

3.2.1.2 Шинэ ойлголт нэмэх

Шинэ ойлголт нэмэх үйл явц 2, 3 болон 4-р алхмуудаар хийгдэнэ. Үүнийг дараах жишээгээр харуулав.

- 2' **НМЦ орчуулагч** өгөгдсөн дэд модонд байж болох *гэзэг* гэдэг ойлголтыг тодорхойлов. Энэ нь *mountain* (уул) гэдэг ойлголттой *нэг-хэсэг* (part-of) утгазүйн холбоосоор холбогдоно. Иймд орчуулагч дараах синсенийг үүсгэж шинэ ойлголт нэмэх саналыг гаргана.

гэзэг -- уул *толгодын ар шил*

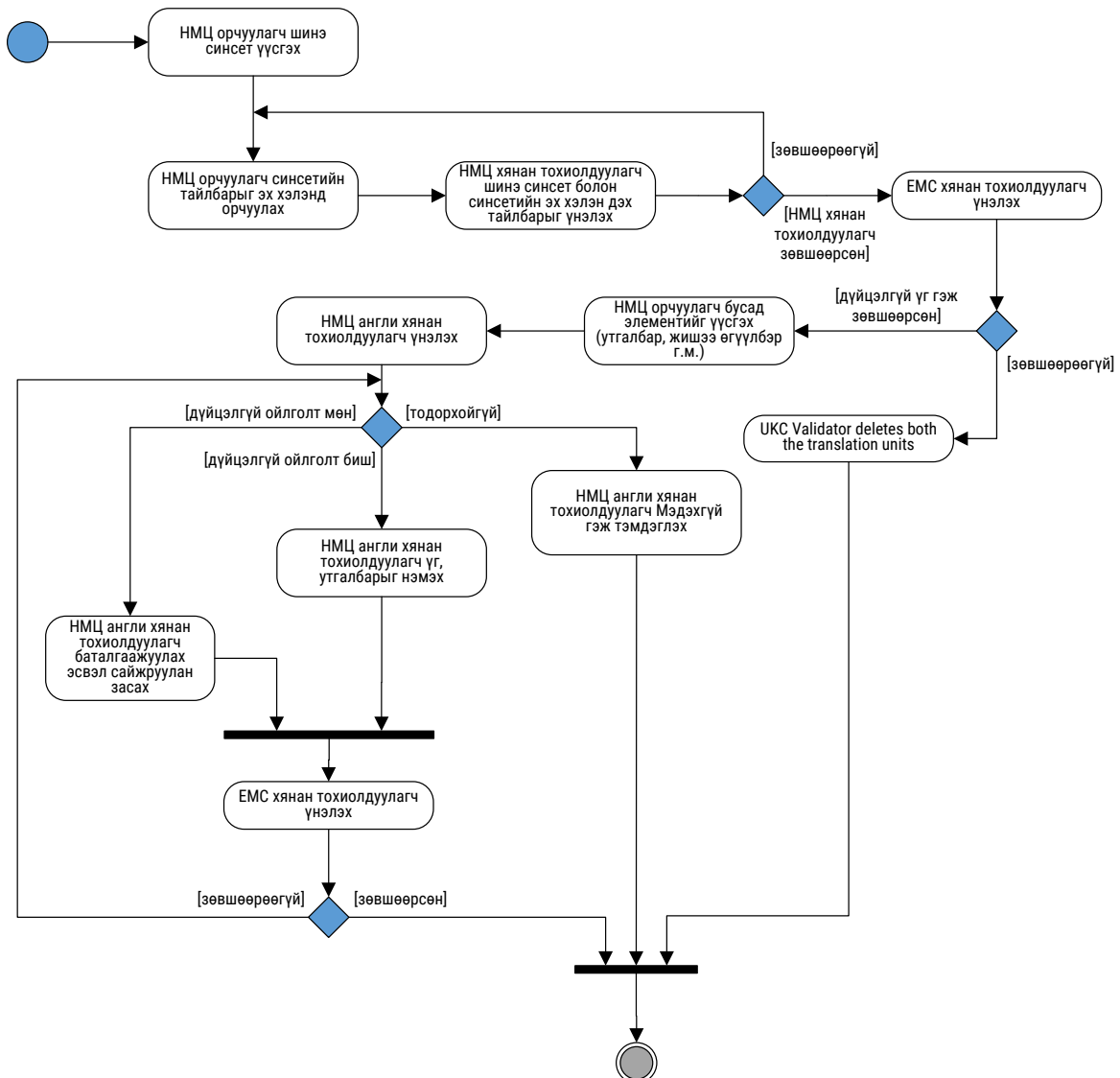
НМЦ англи хянан тохиолдуулагч энэ ойлголтыг англи хэлэнд дүйцэлгүй ойлголт (GAP) болохыг тогтоогоод синсенийн тайлбарыг эх хэл рүү дараах шиг орчуулна.

GAP – *northern ridge of a mountain*

- 3' **EMS Хянан тохиолдуулагч** синсенийн тайлбарын орчуулгыг нягталж англи хэлэнд дүйцэлгүй ойлголт болохыг зөвшөөрнө. Мөн дэд модонд холбогдох

утгазүйн холбоосыг үнэн зөвийг үнэлнэ. Хэлний болон EMC-ийн өнцгөөс энэ ойлголтыг бүрэн шалгасны үндсэн дээр EMC-д оруулахыг EMC Менежерээс асууна.

4' EMC Менежер EMC-д оруулах шаардлагатай ажлуудыг хийж гүйцэтгэнэ.



Зураг 3.9 Шинэ ойлголт нэмэх үйл явцын диаграм

Шинэ ойлголт нэмэх үйл явцын диаграммыг Зураг 3.9-д үзүүлэв.

3.2.2 Мэргэжилтний үүрэг

Хүний түвшний чанарыг гарган авахын тулд мэргэжлийн оролцогчдыг олны хүч болгон зохион байгуулж дараах ангиллыг тодорхойлж тус бүрд нь үүргийг нь тогтоосон (

Хүснэгт 3.3).

Хүснэгт 3.3 Нутгийн мэдлэгийн цөмийн мэргэжилтнүүдийн үүрэг

№	Элемент / Үүрэг	НМЦ орчуулагч	НМЦ хянан тохио-гч*	НМЦ англи хянан тохио-гч*	ЕМС хянан тохио-гч*
А. Үгийн утгалбар					
1	Синсетийн үгс	орчуулах	үнэлэх		үнэлэх
2	Зорилтот хэл дээрх синсетийн үг	нэмэх	үнэлэх		үнэлэх
3	Синсетийн тайлбар	орчуулах	үнэлэх		үнэлэх
4	Синсетийн жишээ өгүүлбэр	орчуулах; нэмэх	үнэлэх		үнэлэх
5	Үглэврийн хувилбар	тодорхойлох	үнэлэх		үнэлэх
6	Үгийн утгалбарын эрэмбэ	тогтоох	үнэлэх		үнэлэх
7	Синсетийн үгийн эрэмбэ	тогтоох	үнэлэх		үнэлэх
8	Утгалбарын товч тайлбар	орчуулах	үнэлэх		үнэлэх
В. Холбоос					
9	Утгалбар / Синсет / Ойлголтын холбоос	тодорхойлох	үнэлэх		үнэлэх
10	Холбоосны төрөл	орчуулах	үнэлэх		үнэлэх
11	Хэлнээс хамаарсан холбоосны шинэ төрөл	тодорхойлох	үнэлэх	орчуулах	үнэлэх
С. Дүйцэлгүй ойлголт					
12	Зорилтот хэлний дүйцэлгүй ойлголт	тэмдэглэх	үнэлэх		үнэлэх
13	Зорилтот хэлний дүйцэлгүй ойлголтын тайлбар	орчуулах	үнэлэх		үнэлэх
14	Эх хэлэнд дүйцэлгүй ойлголт	тодорхойлох	үнэлэх	үнэлэх	үнэлэх
15	Эх хэлний дүйцэлгүй ойлголтын тайлбар	орчуулах	үнэлэх	орчуулах	
16	Эх хэлэнд дүйцэлгүй ойлголтыг илэрхийлэх шинэ үг			тодорхойлох	

* - тохио-гч = тохиолдуулагч

Энд үнэлэх (validate) гэдэг нь аливаа элементийг үнэн зөв болсныг нягталж үзээд хүний түвшний чанарын шаардлага хангана (ассерт) хангахгүй (reject) хэлнэ гэсэн гэсэн үг.

НМЦ орчуулагч

Энэ орчуулагч зорилтот хэлний унаган хэлтэн бөгөөд эх хэлийг сайн эзэмшсэн мэргэжилтэн байна. Мөн тухайн хэл соёлд хэрэглэдэг ерөнхий мэдлэг сайн байх хэрэгтэй. Энэ орчуулагч өөрт оноосон дэд модны доор байрлах хүү ойлголтуудыг утгаар нь ялган харьцуулснаар маш ойлгомжтой орчуулгыг хийнэ. Өөрийн орчуулсан элементийг буцаах (undo) эсвэл устгах эрхтэй байна.

НМЦ хянан тохиолдуулагч

НМЦ хянан тохиолдуулагч нь орчуулагчтай ижил зорилтот хэлний унаган хэлтэн байх бөгөөд хэл шинжлэлийн мэргэжилтэн байна. Зорилгын болон эх хэл шинжлэлийн мэдлэгтэй байх хэрэгтэй. Ийм маягаар тухайн орчуулгыг хэл шинжлэлийн талаас нягталж шалгах юм. Мөн өгсөн үнэлгээг буцаах эсвэл устгах эрхтэй байна.

НМЦ англи хянан тохиолдуулагч

Энэ мэргэжилтэн унаган англи хэлтэн бөгөөд зорилтот хэлийг маш сайн эзэмшсэн байна. Учир нь энэ мэргэжилтэн зорилгын болон эх хэл дээрх дүйцэлгүй ойлголтыг тодорхойлоход хоёр соёлыг сайн мэддэг байх шаардлагатай. Ялангуяа эх хэл дээрх дүйцэлгүй ойлголтыг тодорхойлоход түүний унаган хэлний мэдлэг чухал хэрэгтэй бөгөөд тухайн дүйцэлгүй ойлголт үнэхээр эх хэл соёлд байдаг эсэхийг олж тогтооход илүү дөхөмтэй.

ЕМС хянан тохиолдуулагч

Энэ хянан тохиолдуулагч зорилтот хэлний унаган хэлтэн байх бөгөөд ЕМС-ийн сайн мэдлэгтэй байх ёстой. Ядаж ЕМС-ийн хэлний болон ойлголтын цөмийн ерөнхий ойлголттой байх хэрэгтэй. Энэ хянан тохиолдуулагч НМЦ-ийн менежерийн үүрэг гүйцэтгэх ба НМЦ орчуулагч болон НМЦ хянан тохиолдуулагч нарт орчуулах дэд модыг оноож өгнө. Хэрэв НМЦ орчуулагч эсвэл НМЦ хянан тохиолдуулагч оноосон дэд модыг хүлээн авахаас татгалзвал тэдгээр өөр дэд модыг оноох боломжтой.

ЕМС хянан тохиолдуулагч болон НМЦ англи хянан тохиолдуулагч нар НМЦ орчуулагч болон НМЦ хянан тохиолдуулагч нараас ирсэн дүйцэлгүй ойлголтыг эцэслэн шийдвэрлэх үүрэгтэй. Энэ хоёр хэлнээс хамаарсан ажлууд гүйцэтгэж байхад

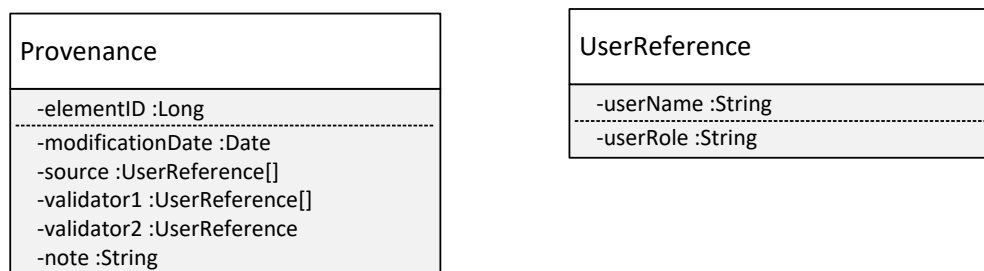
ЕМС менежер ЕМС-ийн мэргэжилтэн байх ба хэлнээс үл хамаарах менежментийн ажлуудыг хийнэ.

3.2.3 Гарал үүслийг зохицуулалт

Аливаа мэдээлэл маш олон төрлийн эх сурвалжаас үүсэж болдог. Гарал үүсэл (provenance) гэдэг бол мэдлэгийн нэгж хэсгийн үүссэн эх сурвалжийг хэлнэ. Энэ ажлын хүрээнд нарийвчилсан гарал үүслийг нутагшуулах элементүүд болон тэднийг үүсгэсэн оролцогчийн гэж тодорхойлсон болно. Эдгээр элементүүдийн үнэн найдвартай зөв байдлыг хангахын тулд гарал үүслийн загварыг ойлголтыг орчуулах болон шинэ ойлголт нэмэх үйл явцад тодорхойлсон [25], [69].

3.2.3.1 Гарал үүслийг илэрхийлэх нь

Гарал үүслийг илэрхийлэх өгөгдлийн бүтцийг Зураг 3.10-т үзүүлэв. Энэ загварын дагуу онтологи нутагшуулах явцад ЕМС-д үүсэж хадгалагдах элементүүдийн гарал үүслийн тухай мэдээллийг хадгалах юм. Эдгээр элементүүдийг Хүснэгт 3.1-д үзүүлсэн.



Зураг 3.10 Гарал үүслийн UML диаграм

Дээрх зурагт үзүүлсэн гарал үүслийн загвар 2 классаас тогтоно. Гарал үүсэл класс нь *<elementID, modificationDate, source, validator1, validator2, note^{op}>* гэсэн 5 шинжтэй. *elementID* гарал үүслийг тодорхойлох элементэд оноосон давтагдахгүй дугаар, *modificationDate* бол гарал үүслийг хамгийн сүүлд өөрчилсөн огноо, *source* ихэвчлэн элементийг орчуулсан эсвэл зорилтот хэл дээр дутуу ойлголтыг санал болгосон оролцогчдын жагсаалт, *validator1* орчуулгыг үнэлсэн оролцогчдын жагсаалт, *validator2* ЕМС хянан тохиолдуулагч, *note^{op}* НМЦ орчуулагч, НМЦ хянан тохиолдуулагч, ЕМС хянан тохиолдуулагч эсвэл ЕМС менежерийн хийсэн (заавал байх албагүй - **optional**) нэмэлт тэмдэглэгээ болно.

UserReference класс *<userName, userRole>* гэсэн 2 шинжтэй ба *userName* оролцогчийн нэр болон и-мэйл хаягийг, *userRole* нь НМЦ орчуулагч, НМЦ хянан тохиолдуулагч болон ЕМС хянан тохиолдуулагч нарын аль нэгийг илэрхийлнэ. Энэ гарал үүслийн

загварт багадаа *source*, *validator1*, *validator2* гэсэн 3 оролцогч холбогдоно. Нутагшуулах элемент бүр 2 хянан тохиолдуулагчаар баталгаажсан байна. Ийм байдлаар элементүүдийн үнэн зөв байдлыг хангана гэж үзсэн.

3.2.3.2 Ойлголт орчуулгын гарал үүсэл

Ойлголт орчуулах макро алхмуудад гарал үүслийг дараах нөхцөл шинэчилнэ.

- a. Хэрэв НМЦ орчуулагч 1(а)-1(б)-р алхам эсвэл 2-р алхмуудад элементүүдийг нутагшуулсан бол элемент бүрийн хувьд шинэ гарал үүсэл *source* шинждээ тухайн орчуулагчийг зааж үүснэ.
- b. Хэрэв НМЦ хянан тохиолдуулагч 1(в)-1(д)-р алхмуудад элементүүдийг үнэлж баталгаажуулсан бол тэдгээр элементүүдийн гарал үүслийн *validator1* шинжид *UserReference* классын тохиолдол болох энэ хянан тохиолдуулагчийг зааж өгнө. Ингэснээр элементүүд үнэлэгдэж баталгаажсан гэх тэмдэглэгээ бий болно.
- c. ЕМС хянан тохиолдуулагч 3-р алхамд элементийг үнэлж баталгаажуулбал тэр элементийн гарал үүслийн *validator2* шинжийн утга энэ хянан тохиолдуулагчаар шинэчлэгдэнэ. Энэ нь тэр элемент бүрэн үнэлэгдэж баталгаажсан, НМЦ-д хүлээн авсан гэх тэмдэглэгээ болно.

Энд *modificationDate* гарал үүсэлд шинээр үйлдэл хийгдэх болгонд шинэчлэгдэж байна. Энэ шинжийн утга (а) үед орчуулга хийсэн огноо, (b) үед орчуулгыг хэлний түвшинд хянан тохиолдуулсан огноо, (c) тохиолдолд НМЦ-д хянан тохиолдуулсан огноо болно.

3.2.3.3 Шинэ ойлголтын гарал үүсэл

Шинэ ойлголт нэмэх үйл явцын үед дараах нөхцөлүүдэд шинэ гарал үүсэл үүсэж шинэчлэгдэнэ.

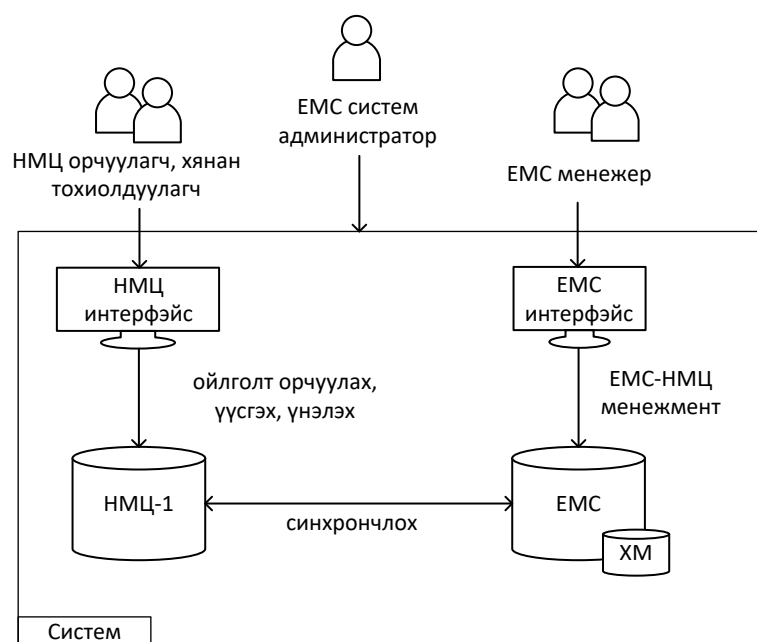
- a. Макро алхмын 2-т НМЦ орчуулагч зорилгын болон эх хэл дээр шинэ ойлголт үүсгэж болно. Энэ тохиолдолд тухайн ойлголттой холбоотой элемент бүрд шинэ гарал үүсэл үүсгэх ба *source* шинжид энэ орчуулагчийг зааж өгнө.
- b. Мөн 2 дахь алхмын үед i) НМЦ англи хянан тохиолдуулагч НМЦ орчуулагчийн үүсгэсэн синсетиийг үнэлж баталгаажуулбал харгалзах гарал үүслийн *validator1* шинжид энэ хянан тохиолдуулагчийг, ii) хэрэв НМЦ англи хянан тохиолдуулагч эх хэл дээр дүйцэлгүй ойлголтыг илэрхийлж болох шинээр үүссэн синсетиийг буцаан зорилтот хэл дээр орчуулбал гарал үүслийн *source*

шинжид энэ хянан тохиолдуулагчийг оноож *validator1* шинжийн утгыг хоосон хэвээр үлдээнэ.

- с. Ойлголтыг орчуулах үйл явцтай ижил EMC хянан тохиолдуулагч шинэ ойлголтыг үнэлж баталгаажуулсан бол гарал үүслийн *validator2* шинжид энэ оролцогчийнг оноож өгнө.

3.2.4 Нутгийн мэдлэгийн цөмийг нутагшуулах систем

НМЦ-ийг нутагшуулах системийн архитектурыг Зураг 3.11-д үзүүлэв.



Зураг 3.11 НМЦ системийн архитектур

НМЦ-д EMC-аас импортолж оруулсан нэг жишиг хэл буюу эх хэл болон ойлголт орчуулах, шинэ ойлголт нэмэх, тэдгээрийг EMC-д экспортлох нэг зорилтот хэлийг хадгалдаг. Үүнд НМЦ орчуулагчид, хянан тохиолдуулагчид НМЦ интерфэйсээр хандаж ажиллана. Системийн бүрэлдэхүүний талаас үзвэл, НМЦ бол тухайн хэл соёлын бүх нутгийн мэдлэгийг хадгалсан EMC-ийн толин тусгал дэд систем болно. Энд хэл соёлоос хамааралтай ойлголтуудыг агуулж байхад EMC нь бүх НМЦ системийн хувьд голд нь байрлаж тасралтгүй мэдээлэл цуглуулж хөгжиж байдаг.

ЕМС нэг жишиг хэлийг, тодруулбал англи хэлийг агуулах ба энэ хэлнээс бүх нутагшуулах үйл явцууд эхлэх юм. EMC-д байгаа ойлголтуудын англи хэлний илэрхийллийг НМЦ-д хуулах ба НМЦ нь хөгжүүлэх зорилтот хэлтэй хамт байна. EMC менежер EMC интерфэйсээр EMC болон НМЦ-ийн менежментийг хариуцан ажилладаг. EMC систем администратор хэрэглэгчийн удирдлагын зохицуулалтыг

хэрэглэгчийн мэдээллийн санд (ХМ) хийж гүйцэтгэх ба системийн засвар үйлчилгээ, арчилгааг бүхэлд нь хариуцна. ЕМС болон НМЦ-ийн үндсэн функцүүдийг доор үзүүлэв.

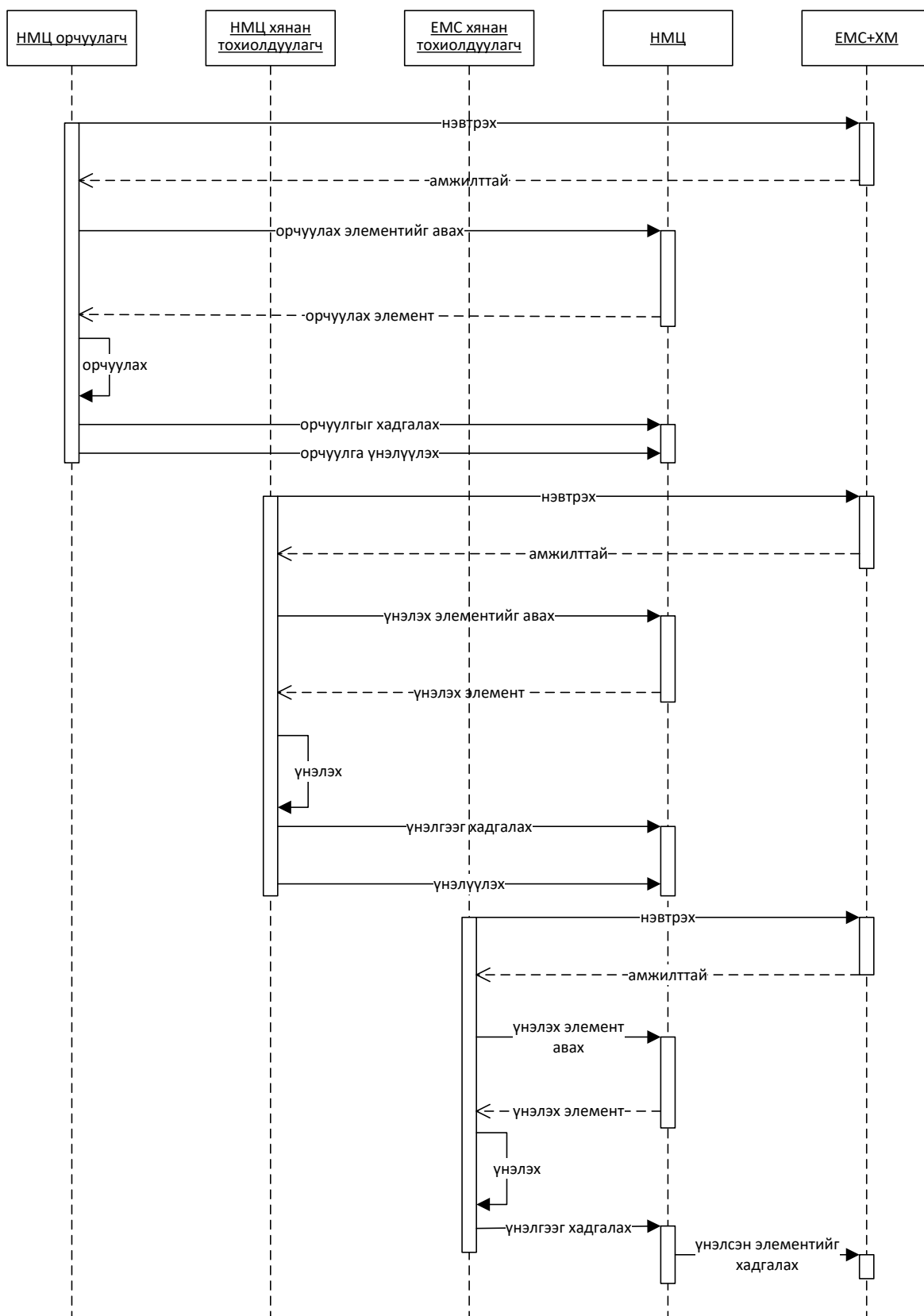
ЕМС-ийн функцүүд

- Жишиг хэлийг ЕМС-ээс экспортолж НМЦ-д хуулах
- НМЦ-д байгаа зорилтот хэлнээс мэдээллийг ЕМС-д импортлох болон нэгтгэх
- Холбогдсон НМЦ, тэдгээрийн хэл, орчуулгын төлөв болон статистик мэдээллийн бүртгэл хөтлөх
- Хэрэглэгчийн бүртгэл болон хэрэглэгчийн хандах эрхийн зохицуулалтыг Хэрэглэгчийн санд гүйцэтгэх; нэг Хэрэглэгчийн сан бүх НМЦ системд ашиглана
- Хэрэглэгчийн үүргийн тодорхойлолт болон НМЦ-ийн орчуулах болон хянан тохиолдуулах даалгаврын оноолтыг хийх;

НМЦ-ийн функцүүд

- Нутагшуулах дэд модыг боломжит оролцогчид оноох
- НМЦ-ийн зорилтот хэлэнд элементүүдийг орчуулга хийх
- НМЦ-ийн зорилтот хэлэнд элементүүдийг хянан тохиолдуулах
- Шинэ ойлголт нэмэх (ЕМС-д шинэ ойлголт эсвэл орхигдсон ойлголт)
- Гадаад эх сурвалжаас мэдээллийг импортлох
- Зорилтот хэл дээр байгаа элементүүдийг экспортлох

НМЦ-ийн элементүүдийг орчуулах болон хянан тохиолдуулах дарааллыг диаграммыг жишээ болгон Зураг 3.12-т үзүүлэв. Энэ диаграмд НМЦ орчуулагч болон хянан тохиолдуулагч нар нутагшуулах элементийг НМЦ-ээс авч НМЦ-д буцаан хадгалах бол ЕМС хянан тохиолдуулагч ЕМС-д хянан тохиолдуулсан элементийг хадгална. Системийн талаас харвал, НМЦ-ийн элемент (Хүснэгт 3.1-д үзүүлсэн үглэвэр, синсет, утгалбар г.м.) нэг бүрийн хувьд орчуулах, үнэлэх болон хадгалах үйлдэл хийнэ.



Зураг 3.12 Орчуулах болон үнэлэх үйл явцын дарааллын диаграм

KOS Lexicalization Knowledge

oyundari.nyamdavaa Logout

Localization Application [Back to dashboard](#)

Reference Language: English [Provenance](#)

Conceptid: 66269

Gloss: an empty area (usually bounded in some way between things)

POS: NOUN

Senses	Rank	Lemma	Exceptional forms
	1	space	

Examples: Missing Example

Target Language: Mongolian [LOG](#) [Translate Next](#)

Synset

Is GAP? [Signal as GAP](#)

Gloss: хоосон орон зай (ихэвчлэн эргэн тойрондоо ямарваа зүйлээр хүрээлэгдсэн)

POS: Noun

Senses

Rank	Lemma	Exceptional word forms
1	хоосон зай	Exceptional form... + ✎
2	зай завсар	Exceptional form... + ✎
3	огторгуу	Exceptional form... + ✎

Add sense (write I) Exceptional forms for the lemma, comma separated [+ Add](#)

Examples

Add Example [+ Add](#)

[Save](#) [Send to validation](#)

Зураг 3.13 LKC-UI-ийн НМЦ орчуулагч хэрэглэгчийн интерфэйс

Энэ диаграмд орчуулах гэдэг нь синсенийн үгийг орчуулах, шинээр үг нэмэх, синсенийн үгсийн эрэмбэ тогтоох гэх мэт (Хүснэгт 3.3) бүх үйлдлийг илэрхийлнэ.

Жишээ нь, тухайн үгийн хувьд утгалбар шинээр үүсгэнэ гэдэг бол орчуулах гэдэг нэг үйлдлээр илэрхийлэгдэж байгаа болно.

НМЦ орчуулагчийн синсет орчуулах интерфэйсийг Зураг 3.13-д үзүүлэв. Энэ зурагт НМЦ орчуулагч эх хэл буюу жишиг хэл (reference language) дээр байгаа синсенийн үг *space*, түүний тайлбар *an empty area (usually bounded in some way between things)* гэсэн тайлбарыг зорилтот хэлэнд (target language) орчуулсан байна. Ингэхдээ *хоосон зай*, *зай завсар*, *огторгуу* (үг үсгийн алдаатай) гэсэн синсенийн 3 үгсийг эрэмбэлэн синсенийн тайлбарыг *хоосон орон зай (ихэвчлэн эргэн тойрондоо ямарваа зүйлээс хүрээлэгдсэн)* гэж орчуулсан байна. *хоосон зай* гэдэг үгийн нэг утгалбар нь энэ синсенийн утга болно. Энэ үг өөр утгаараа өөр үгстэй нийлж өөр синсет үүсгэж болох тул *Senses* хэсгийн дор оруулж өгсөн учиртай. Бусад интерфэйсийг Хавсралт Г. НМЦ-ийг нутагшуулах системийн хэрэглэгчийн зарим интерфэйсээс үзнэ үү. Энэ програмыг Италийн Трэнтогийн их сургуулийн KnowDive судалгааны багийн програм зохиогчид програмчлалын Жава хэл, өгөгдлийн сангийн PostgreSQL менежмент системийг ашиглан хөгжүүлсэн болно.

Example

Target

Synset

Is GA

Gloss

POS

Sense

Rank

1

зай

Exceptional form...

LOG Tr

хүрээл

+

+

Log for synset 145614

	Type	Status	Comment
Jun 16, 2016 4:34:08 by altangerel.chagnaa	SYNSET	VALIDATED_LKC	
Dec 24, 2014 8:28:10 by enkhmaa.zorigt	SYNSET	VALIDATION_LKC_PENDING	Submitting translation for LKC validation
Dec 24, 2014 8:28:10 by enkhmaa.zorigt	SYNSET	TRANSLATED	Synset translated

1

Close

Зураг 3.14 LKC-UI-ийн гарал үүслийн бүртгэлийн цонх

Хэрэв энэ нь дүйцэлгүй үг байсан бол *Signal as GAP* товчийг дарж холбогдох тайлбарыг орчуулах өөр интерфэйс рүү шилжинэ. НМЦ орчуулагч *Save* товчийг дарж орчуулгыг хадгалах боломжтой. Мөн дараа нь эргэн орж ирж өөрчлөлт хийх эсвэл орчуулгыг болсон гэж үзвэл *Send to validation* товчийг дарж НМЦ хянан

тохиолдуулагч руу явуулах боломжтой. Хэрэв НМЦ хянан тохиолдуулагч энэ орчуулгын зөвхөн синсетийг үнэлж хүлээн зөвшөөрсөн бол энэ үйл явцын гарал үүслийг Зураг 3.14-д үзүүлэв. Энэ мэтчилэн бусад элемент нэг бүрийн хувьд гарал үүслийн бүртгэлийг хөтөлнө.

Гадаад эх сурвалжаас мэдээллийг импортолж оруулах файлын загварыг хүснэгтэн хэлбэрээр (spreadsheet эсвэл CSV – Comma Separated Values) боловсруулсан. Энэ файлын загварын дагуу гар аргаар орчуулга гүйцэтгэж болох бөгөөд НМЦ-д шууд импортлон оруулах боломжтой. Мөн RDF загвараар НМЦ систем онтологийг экспортлох боломжтой.

3.2.4.1 Файлын хүснэгтэн загвар

Энэ файлын загвар 3 хүснэгт талбараас (sheet) тогтоно. Эхнийх нь *senses*, үгийн утгалбар болон синсетийн тайлбарыг, удаах нь *relations*, ойлголтуудын хоорондох холбоосуудыг, сүүлийнх нь *gaps*, зорилгын болон эх хэл дээрх дүйцэлгүй ойлголтыг тус тус агуулна.

senses дараах багануудтай:

- **Cased word lemma**

Том жижиг үсгийн ялгаатай үгийн сангийн нэгж

- **Word forms**

Цэг таслалаар заагласан үглэврийн хувилбар (байвал)

- **Concept UK ID**

Ойлголтын дугаар

- **Word Sense Rank**

Тухайн хэлний үгийн санд байх үгийн утгалбарын эрэмбэ

- **Concept Word Rank**

Тухайн ойлголттой холбогдсон синсетийн үгийн эрэмбэ

- **PoS**

Синсетийн үгийн аймгийн тэмдэглэгээ (нэр, үйл, тэмдэг нэр болон дайвар үгийн аль нэг)

- **Description**

Синсетийн тайлбар (жишээ өгүүлбэр байвал цэг таслалаар зааглаж бичих)

- **Operation**

Хэрэглэгчийн хийх үйлдэл: ADD-нэмэх, UPDATE-шинэчлэх, DELETE-устгах

- **Language**

Хэлний код (ISO 639-1:2002), жишээ нь, en, it, mn гэх мэт.

Шинэ ойлголтууд хасах тоогоор дугаарлагдана.

relations дараах багануудтай:

- **Parent Concept UK ID**

Холбоосонд оролцох эцэг ойлголтын дугаар

- **Child Concept UK ID**

Холбоосонд оролцох хүү ойлголтын дугаар

- **Relation Kind**

Эцэг болон хүү ойлголтын хоорондох холбоосны төрөл. Жишээ нь, IS_A болон PART_OF.

- **Operation**

Хэрэглэгчийн хийх үйлдэл: ADD-нэмэх, UPDATE-шинэчлэх, DELETE-устгах

- **Parent**

Эцэг ойлголтыг төлөөлөх үглэвэр

- **Child**

Хүү ойлголтыг төлөөлөх үглэвэр

Сүүлийн хоёр талбар, Parent болон Child, системд импортлоход ашиглагдахгүй, хүмүүс хялбар ойлгоход туслах үүрэгтэй талбар юм.

gaps дараах багануудтай:

- **Concept UK ID**

Ойлголтын дугаар

- **Language**

Хэлний код (ISO 639-1:2002) жишээ нь, en, it, mn гэх мэт.

- **Operation**

Хэрэглэгчийн хийх үйлдэл: ADD-нэмэх, UPDATE-шинэчлэх, DELETE-устгах

Гарал үүслийн мэдээллийг хадгалахад зориулж бүх хүснэгтэн талбаруудад дараах 3 баганууд нэмэгдэнэ.

- **Reference**

Үйлдлийг гүйцэтгэсэн хэрэглэгчийн нэр, и-мэйл хаяг

- **Note**

Нэмэлт тэмдэглэгээ

- **Reference_Type**

Хэрэглэгчийн төрөл

Энэ загварыг Хавсралт Б. Файлын хүснэгтэн загварын жишээгээр дэлгэрэнгүй үзүүлэв.

3.2.4.2 RDF загвар

Хүснэгт 3.4-т файлын хүснэгтэн болон RDF загварт хэрэглэх элементүүд хоорондын буулгалтыг үзүүлэв. RDF загварыг ВөрдНэт RDF²⁰ стандартад суурилж хөгжүүлсэн. Энэ хүснэгтэд байгаа *wn20schema*: нэрлэлтийг (namespace) ВөрдНэт 2.0 RDF схемээс²¹ харж болно. Харин EMC-ийн элементүүдэд ашиглах *uk*: нэрлэлтийг²² нэмж хөгжүүлсэн.

Хүснэгт 3.4 RDF болон файлын хүснэгтэн загвар хоорондох буулгалтын жишээ

Хүснэгтэн загварын элементүүд	RDF элементүүд
<i>Senses</i>	
Cased Word Lemma	<code>wn20schema:senseLabel</code>
Word Forms	<code>uk:wordForm</code>
Concept UK ID	<code>uk: conceptUKID</code>
Word Sense Rank	<code>uk:wordSenseRank</code>
Concept Word Rank	<code>uk:synsetWordRank</code>
PoS	<code>uk:partOfSpeech</code>
Description	<code>wn20schema:gloss</code>
Operation	<code>uk:operation</code>
Language	<code>xml:lang</code>
<i>Relations</i>	
Parent Concept UK ID	<code>uk:parentConceptUKID</code>
Child Concept UK ID	<code>uk:childConceptUKID</code>

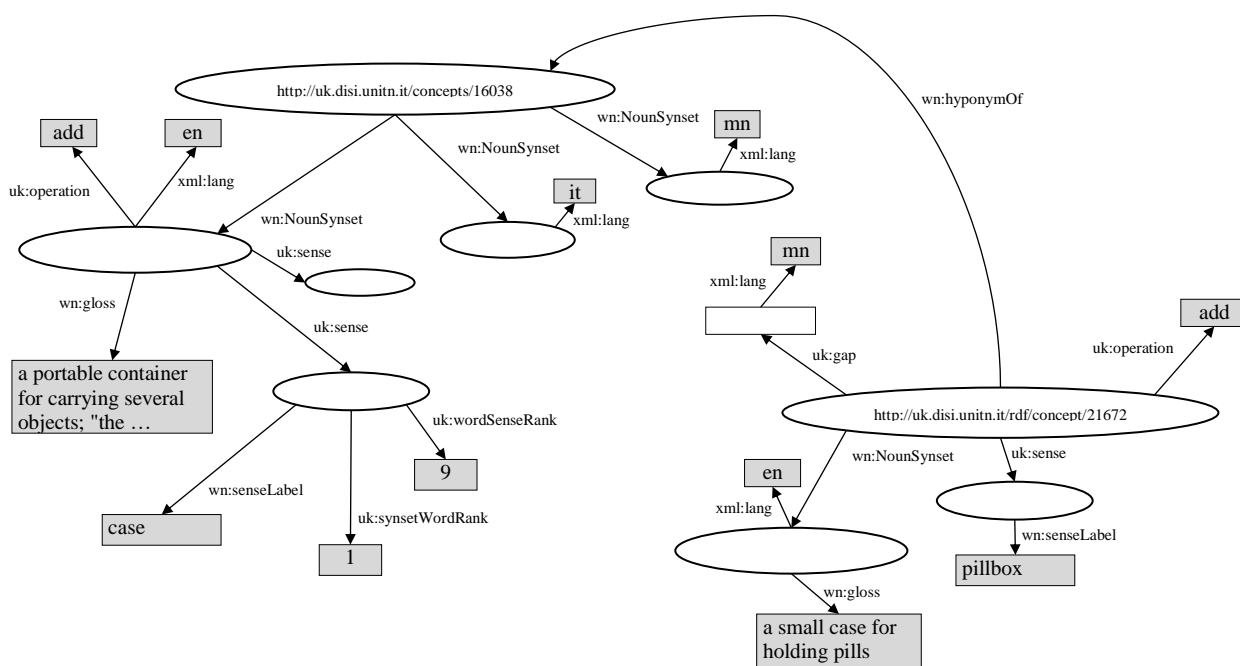
²⁰ <http://www.w3.org/TR/wordnet-rdf>

²¹ <http://www.w3.org/2006/03/wn/wn20/schema>

²² <http://ukc.kidf.eu>

Relation Kind	owl:ObjectProperty
Operation	uk:operation
Parent	wn20schema:senseLabel
Child	wn20schema:senseLabel
IS_A	wn20schema:hyponymOf
PART_MERONYM	wn20schema:partMeronymOf
<i>Gaps</i>	
Concept UK ID	uk:conceptUKID
Language	xml:lang
Operation	uk:operation

Хүснэгт 3.4-д үзүүлсэн буулгалтын дагуу, хүснэгтэн загвараар хөгжүүлсэн жишээ өгөгдлийг RDF загварт буулгасныг Зураг 3.15-д үзүүлэв.

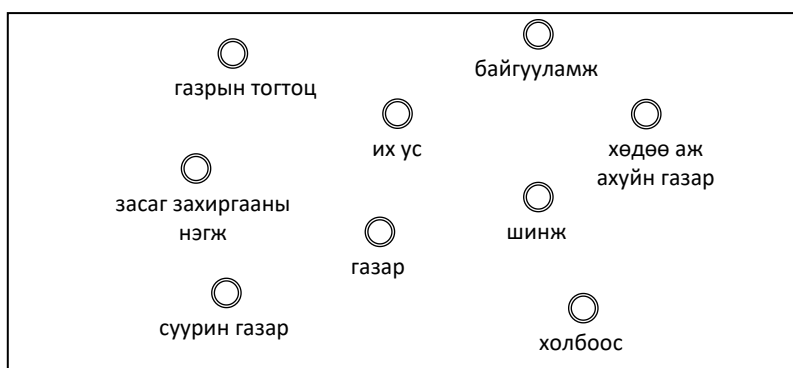


Зураг 3.15 Хүснэгтэн загвараар хөгжүүлсэн өгөгдлийг RDF-д буулгасан жишээ

wn нэрлэлт бол wn20schema нэрлэлтийн товчилсон хувилбар болно. Энэ зурагт case гэсэн төлөөлөх үглэвэртэй 16038 дугаартай ойлголт англи, итали, монгол зэрэг хэл дээрх 3 синсеттэй холбоотойг харуулж байна. Англи синсет ганц үгтэй бөгөөд түүний синсенийн үгийн эрэмбэ нь 1, үгийн утгалбарын эрэмбэ нь 9 болно. 21672 дугаартай ойлголт нь Монгол хэлний хувьд дүйцэлгүй ойлголт байна. Үүнийг хэрэгжүүлсэн RDF баримтыг Хавсралт В. Хэлний цөмийг экспортлох RDF файлын жишээнээс үзнэ үү.

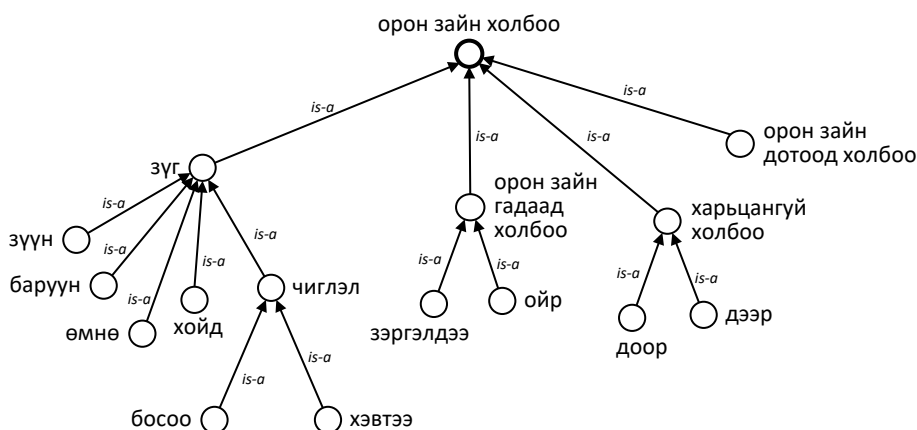
3.2.5 Туршилт, үр дүн

Энэ ажлаар Орон зайн айн [41] онтологийг нутагшуулах туршилт хийсэн. Орон зайн айн бол онтологийг хөгжүүлэх фасетэд аргачлалаар [42] хөгжүүлсэн газар зүйн маш том онтологи юм. Энэ онтологийг GeoNames болон ВөрдНэт нэгтгэж үүсгэсэн. Үүнийг орон зайн онтологи гэж нэрлэж болно. Энэ онтологи 17 фасет, 985 ойлголт болон 8.5 сая нэгж объектуудаас тогтдог бөгөөд EMC-д орсон болно. Зарим фасетуудыг дурдвал, *газрын тогтоц* - *land formation* (уул, толгод г.м.), *их ус* - *body of water* (далай, нуур г.м.), *байгууламж* - *facility* (их сургууль, үйлдвэр г.м.) зэрэг болно.



Зураг 3.16 Орон зайн айн фасетуудын дэд хэсэг

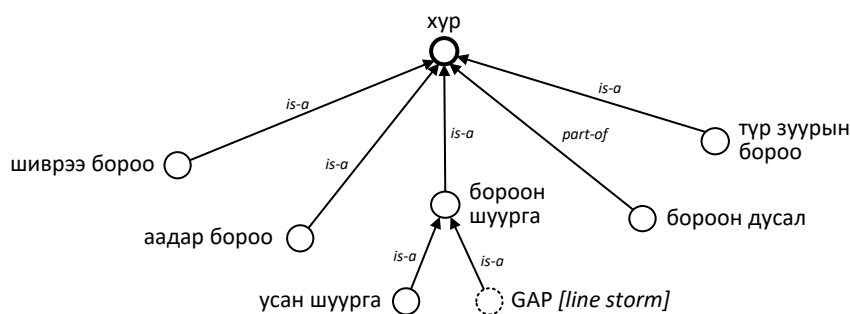
Зураг 3.16-д Орон зайн айн фасетуудын хэсэгчлэн үзүүлэв. Энд эдгээр фасетууд нь өөр хоорондоо холбоогүй бөгөөд ойлголтууд нь фасет дотор болон фасетуудын дунд давхацдаггүй. EMC-ийн нэгж-объектын класс фасетийн жишээг Зураг 3.7-д газарзүйн тогтоц фасетийн дэд хэсгээр үзүүлсэн.



Зураг 3.17 Орон зайн холбоос (R) фасетийн дэд хэсэг

Орон зайн айд *холбоос (R)* төрлийн *орон зайн холбоос (spatial relation)*, *эх урсгал (primary outflow)* зэрэг 10 гаруй фасет байдаг. Орон зайн холбоос фасетийн зарим хэсгийг Зураг 3.17-д үзүүлэв. Энэ холбоосны фасет нь газарзүйн хоёр физик объектын

хоорондох эсвэл ямар зүйлийн байрлалыг заах орон зайн шинж чанар юм. Жишээ нь, Монгол улс ОХУ-ын *өмнө* (south), БНХАУ-ын *хойд* (north) талд оршдог. Харин эх урсгал бол хоёр *их усны* (body of water) хоорондох холбоос юм. Дээрх хоёр зурагт давхар зураастай дугуй дүрсээр эх ойлголтыг (root concept) бусад хүү ойлголтуудаас ялгаж тэмдэглэсэн болно. Мөн энэ айд *шинж* (A) төрлийн *хур*, *температур* гэх мэт 20 орчим фасет байдаг.



Зураг 3.18 хур шинжийн (A) фасетийн дэд хэсэг

Зураг 3.18-д үзүүлсэн хур фасет нь агаар мандалд хий төлөвөөс шингэн төлөвт орсон уур ус болон дуslaх гэсэн ойлголт болно. *температур* нь хүрээлэн буй орчин эсвэл биеийн (түүний молекулын ажиллагаанаас хамаарах) халуун эсвэл хүйтний зэрэг юм.

Энэ ажлаар орон зайн онтологийн нийт ойлголтын 95.7%-ийг Монгол хэл рүү нутагшуулсан бөгөөд үлдсэн 4.3% нь дүйцэлгүй ойлголт байсан. Бодит жишээг Хавсралт Е. Дүйцэлгүй ойлголтын жишээнээс дэлгэрэнгүй үзнэ үү. Хүснэгт 3.5-д нутагшуулалтын статистик болон хүрсэн үр дүнг үзүүлэв. Энэ хүснэгтийн мөрөнд фасет бүрийн ойлголтыг тоог харуулав. Жишээ нь, засаг захиргааны нэгж (administrative division) 18 ойлголттой, хөдөө аж ахуйн газар (agricultural land) 19 ойлголттой гэх мэт. Энд бүх шинж болон холбоос фасетуудыг *attribute* болон *relation* гэж бүлэглэсэн болно. НМЦ орчуулагчид эхний орчуулгаар 910 ойлголт орчуулж, 75 дүйцэлгүй ойлголтыг зорилтот хэлэнд (Монгол) тодорхойлсон. НМЦ хянан тохиолдуулагч болон ЕМС хянан тохиолдуулагч нар синсетийн үгс болон синсетийн тайлбарт бүрийг үнэлж санал өгсөн. Үнэлгээний явцад синсетийн үгсийн хувьд 611 санал зөрсөн тохиолдол, синсетийн тайлбарын хувьд 769 тохиолдол гарсан. Санал зөрсөн, орчуулгад өөрчлөлт оруулах нь орчуулагч болон хянан тохиолдуулагч нар санал нийлэх хүртэл давтагдана. Хамгийн их давтагдсан тоо 4 байсан.

Хүснэгт 3.5 Орон зайн онтологийг нутагшуулсан үр дүн

Фасет	Ойлголт	Үг	Орчуулгын үр дүн				Үнэлгээний үр дүн				Нутагшуулсан ойлголт
			Ойлголт	Үг	Санал болгосон дүйцэлгүй ойлголт		Санал зөрсөн үг	Санал зөрсөн синсегийн тайлбар	Зөвшөөрсөн дүйцэлгүй ойлголт		
					Монгол хэлэнд	Англи хэлэнд			Монгол хэлэнд	Англи хэлэнд	
administrative division	18	30	18	35	0	0	12	4	0	1	19
agricultural land	19	28	19	26	0	7	7	9	0	7	26
attribute	85	120	75	148	10	0	58	72	8	0	77
barren land	7	8	7	9	0	1	4	6	0	1	8
facility	357	616	355	421	2	34	197	261	2	36	391
forest	5	6	5	7	0	2	6	6	0	2	7
geological formation	200	273	152	243	48	31	151	186	26	31	205
land	15	17	13	18	2	2	11	15	0	1	16
plain	12	16	9	15	3	2	2	3	2	2	12
rangeland	8	14	8	14	0	0	8	13	0	1	9
region	46	62	44	50	2	0	22	12	0	0	46
relation	54	111	54	120	0	0	44	77	0	0	54
wetland	8	13	8	18	0	8	5	5	0	8	16
abandoned facility	16	16	15	16	1	0	13	22	0	0	16
body of water	116	146	113	130	3	8	58	69	2	8	122
populated place	13	16	11	21	2	1	9	6	2	1	12
seat of government	6	9	4	6	2	0	4	3	0	0	6
Нийт элемент	985	1501	910	1297	75	96	611	769	42	99	1042

НМЦ хянан тохиолдуулагчид НМЦ орчуулагчдын тодорхойлсон нийт 75 дүйцэлгүй ойлголтоос 43 худал нь үнэн (false positive) дүйцэлгүй ойлголтыг илрүүлсэн. Эцэст нь 42 дүйцэлгүй ойлголтыг тодорхойлж нийт англи хэлний 985 ойлголтоос 943 ойлголтыг нутагшуулсан. НМЦ орчуулагчдын санал болгосон эх хэлний 96 дүйцэлгүй ойлголтыг (шинэ эсвэл орхигдсон) НМЦ хянан тохиолдуулагчид 99 болгон баталгаажуулжээ. Энэ нь онтологи нутагшуулалтын явцад НМЦ хянан тохиолдуулагчид дүйцэлгүй ойлголтыг орчуулагчид санал болгож байсныг илтгэнэ.

3.2.5.1 Ялгамжийн томъёолол

Ялгамжийн олон төрлүүдээс дүйцэлгүй ойлголтыг томъёолж болно. Ай нь фасетуудаас тогтох бөгөөд фасет нь төрөл ойлголтуудыг агуулна. Жишээлбэл, Хүснэгт 3.5-д үзүүлсэн үр дүнгээс харахад англи хэлээр илэрхийлсэн онтологийн газарзүйн тогтоц (geological formation) фасет нь 200 ойлголттой. Эндээс бид эх хэлээр илэрхийлэх онтологи, V_s векторыг $\{f_1, f_2, f_3 \dots f_n\}$ гэх мэтчилэн f_n фасетуудын олонлог гэж үзэж болно. Тэгвэл онтологи нутагшуулалтын үр дүнд үүсэх зорилтот хэлээр илэрхийлэх онтологийг ойлголтуудыг агуулах V_t векторыг (3.1) томъёогоор илэрхийлж болно.

$$V_t = V_s + (D_{st} - D_{ts}) \quad (3.1)$$

Энд D_{st} бол эх хэлнээс зорилтот хэлэнд тодорхойлсон дүйцэлгүй ойлголтыг илэрхийлэх вектор, D_{ts} нь зорилтот хэлнээс эх хэлэнд тодорхойлсон дүйцэлгүй ойлголтыг илэрхийлэх вектор болно.

3.3 Вэб хэрэглэгчдээр онтологи нутагшуулах аргачлал

Өмнөх дэд бүлэгт онтологийг мэргэжлийн олны хүчээр нутагшуулах аргачлалыг боловсруулж туршсан билээ. Гэвч мэргэжлийн олны хүчийг зохион байгуулах нь энгийн вэб хэрэглэгчдийг зохион байгуулахаас илүү хүнд байх нь ойлгомжтой. Учир нь нэг талаас дээрх аргачлалд хэл шинжээч (НМЦ хянан тохиолдуулагч), эх болон зорилтот хэл соёлыг сайн мэддэг англи хэлний мэргэжилтэн (НМЦ англи хянан тохиолдуулагч), онтологийн инженер (ЕМС хянан тохиолдуулагч, ЕМС менежер) зэрэг мэргэжлийн оролцогчдыг шаардана. Нөгөө талаас бага үнэлэмжтэй оролцогчдыг бодвол илүү өндөр зардалтай. Хүснэгт 3.2-д харуулсан олны хүчээр хийх ажлуудад шаардлагатай оролцогчдын үнэлэмжийг тодорхойлсон. Энэ ажлуудаас синсетийн үгсийг орчуулах, синсетийн үгсийн эрэмбийг тогтоох, зорилтот хэлэнд дүйцэлгүй үгийг тодорхойлох зэрэг бага үнэлэмжтэй оролцогчдоор хийлгэж болох ажлыг энгийн

вэб хэрэглэгчдээр гүйцэтгүүлэх олны хүчний аргачлалыг боловсруулсан. Гол санаа нь бага үнэлэмжтэй оролцогчдын хийж чадах ажлыг вэб хэрэглэгчдээр гүйцэтгүүлээд үлдсэн ажлыг нь мэргэжлийн олны хүчээр гүйцэтгүүлэх юм. Вэб хэрэглэгч гэдгийг интернетэд холбогдсон компьютерийн ард сууж ажиллах хэн нэгнийг гэж ойлгоно. Энэ оролцогч ямар мэдлэг туршлагатай, хэдэн настай хүн гэдгийг бид мэдэхгүй. Энэ аргачлалын зорилго нь олон тооны вэб хэрэглэгчдээр гүйцэтгүүлэх ХОД боловсруулж эх хэлний синсетэд хамгийн сайн тохирох зорилтот хэлний синсетийг олох юм.

3.3.1 Хүний оюуны даалгаврын зохиомж

Энэ ажлаар 2 үет олны хүчээр синсетийг орчуулах аргачлалыг боловсруулсан болно. Эхний үе нь синсетийг орчуулах, хоёр дахь нь синсетийн үгсийг үнэлэх юм (Зураг 3.19).

3.3.1.1 Синсет орчуулах

Энэ аргачлалын хувьд синсет орчуулах гэдэг нь эх хэлний синсетийн үгсийг зорилтот хэлэнд орчуулах боломжтой бол зорилтот хэлний синсетэд үг нэмэх ажлыг хэлнэ. Өөрөөр хэлбэл, эх хэлний синсетэд хамгийн сайн тохирох зорилтот хэлний синсетийг үүсгэх буюу түүний үгсийг оноох юм.

Синсет орчуулах даалгаврын зохиомжид тавих шаардлага

1. Вэб хэрэглэгчид үгийг орчуулахдаа синсетийн англи үгийг харьцуулж харах шаардлагатай. Энэ нь тухайн синсетийг илэрхийлж чадах Монгол үгийг онооход хэрэглэгчид сэдэл төрүүлэх боломжтой байх ёстой. Жишээ нь: {reservoir, artificial lake, man-made lake} синсетийн үгсээс artificial lake үг хэрэглэгчид сэдэл төрүүлж ядаж нэг үг оноох боломжийг гарган өгч болно.
2. Вэб хэрэглэгч үг оноохдоо англи синсетийн илэрхийлэх ойлголтыг сайн ойлгосон байх шаардлагатай бөгөөд синсетийн тайлбарыг (боломжтой бол бусад илэрхийлэх материалыг, жишээ нь, жишээ өгүүлбэр, зураг) харуулах шаардлагатай.

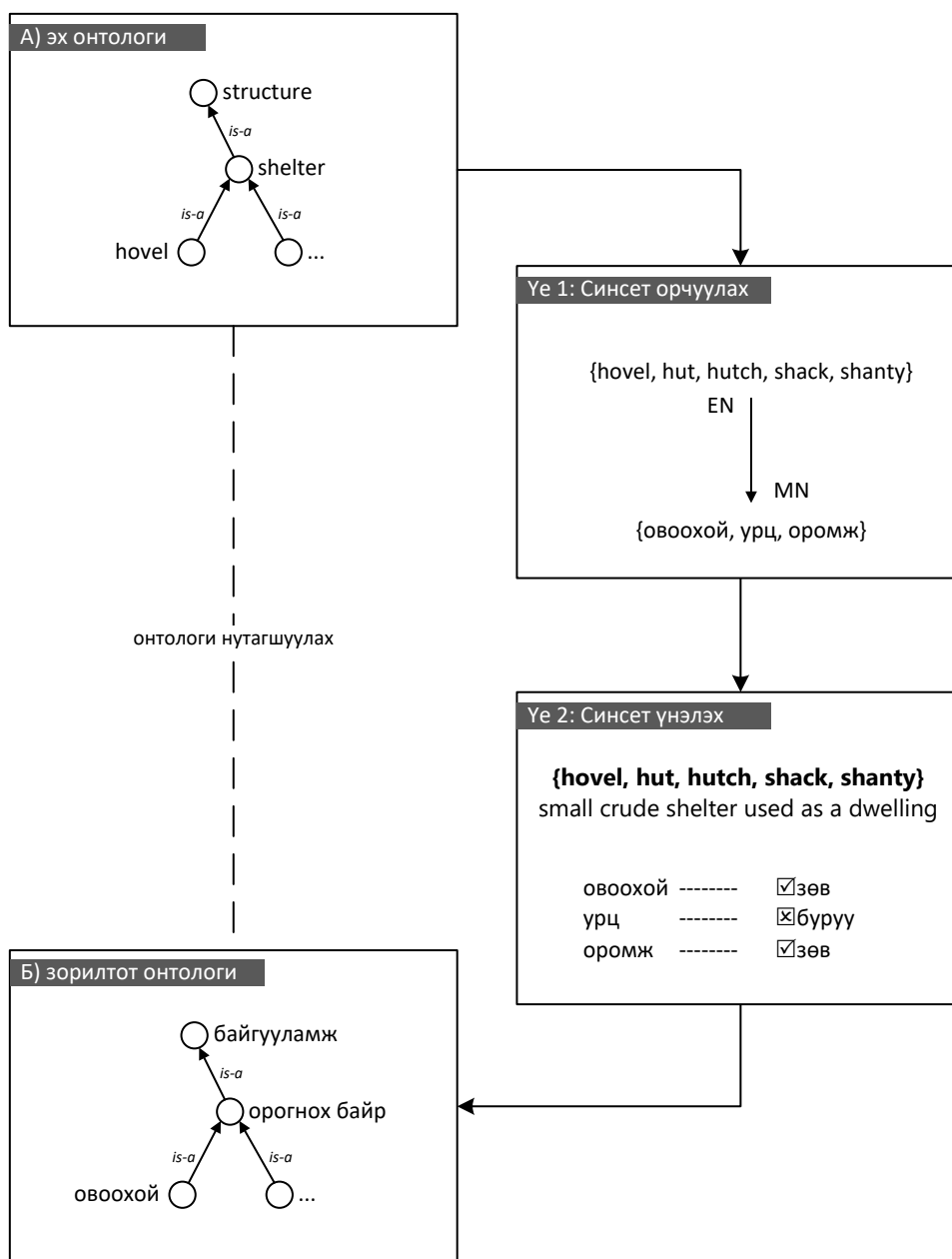
Вэб хэрэглэгчдээс дараах асуулт, зааврын дагуу ажлыг гүйцэтгэхийг асуусан.

Даалгаврын асуулт: Англи хэлээр илэрхийлсэн дараах ойлголтод тохирох Монгол үгийг оноож бичнэ үү.

Заавар: Орчуулга хийхдээ дараах зааврын дагуу хийнэ.

Үг оноох

Эхлээд англи хэлээр илэрхийлсэн ойлголтыг илэрхийлэх үгс, ойлголтын тайлбар, жишээ өгүүлбэр зэргийг сайтар ойлгож дараа нь тэр ойлголтыг илэрхийлэх Монгол үгийг оноож бичнэ.



Зураг 3.19 Вэб хэрэглэгчдээр онтологи нутагшуулах аргачлалын бүдүүвч

Оноох үг нь үг сангийн НЭГ нэгж байна. Энэ бол толь бичгийн НЭГ толгой үг эсвэл НЭГ хэлц үг эсвэл НЭГ холбоо үг юм. Жишээ нь:

1. толгой үг: толь бичгийн нэг зүйл үг; жишээ нь: буудал; зам г.м.
2. хэлц үг: нэг утга санааг илэрхийлэх, аль нэг үгийг өөр үгээр сольж огт болохгүй нийлэмж үг; жишээ нь: цэнхэр дэлгэц;

3. холбоо үг: хоёр ба түүнээс дээш үг өгүүлбэр зүйн ямар нэг холбоогоор холбогдож, нэрлэх үүрэгтэй нийлмэл үг; жишээ нь: онгоцны буудал;

Жишээ нь:

Англи ойлголт

store; shop

a mercantile establishment for the retail

sale of goods or services

Монгол ойлголт

дэлгүүр

Дүйцэлгүй ойлголт

Хэрэв тухайн ойлголтыг Монгол хэлэнд илэрхийлэх боломжгүй бол түүнийг дүйцэлгүй ойлголт гэнэ. Өөрөөр хэлбэл үгийн сангийн нэгжээр илэрхийлэх боломжгүй ба олон чөлөөт үгээр илэрхийлэхээр байвал GAP! гэсэн товчийг дарж тэмдэглэнэ үү.

Анхаарах зүйлс

Үг оноож бичихэд

1. Зөвхөн КИРИЛЛ үсгээр зөв бичих дүрмийн алдаагүй байх
2. Үгийн сангийн нэгж (толгой үг | хэлц | холбоо үг) байх
3. Том жижиг үсэг оролцуулж болно. Жишээ нь: Христийн сүм
4. Хоёроос дээш үг оноож бичих тохиолдолд цэг таслалаар (;) зааглаж бичих

Дараах нөхцөлд таны орчуулгыг хүлээж авахгүй

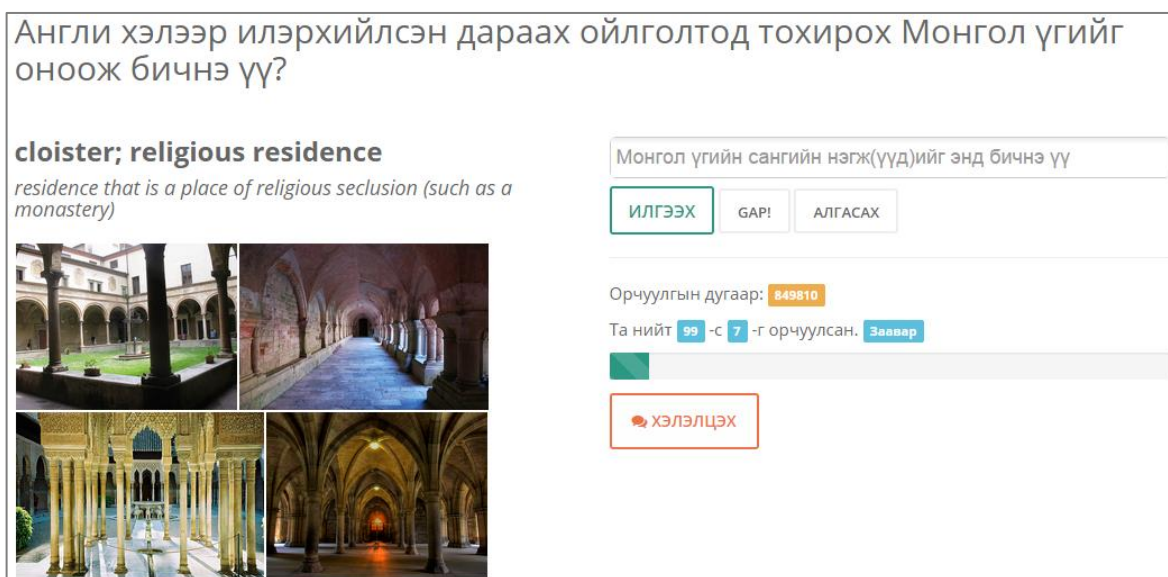
1. Бичвэр оруулахдаа латин үсэг эсвэл үг эсвэл цифр оролцуулсан бол,
жишээ нь: [zuslan]; [vзэг] - ‘ү’ үсгийн оронд англи ‘v’ үсэг; [GAP]; [2 тийшээ салсан тохой]
2. Монгол хэлэнд хэрэглэгддэг - (дундуур зураас) ба ; (цэг таслал)-аас бусад цэг цэглэлийн тэмдэгтүүдийг хэрэглэсэн бол,
жишээ нь: . (цэг), : (тодорхойлох цэг), () (хаалт), / (налуу зураас) гэх мэт.

Зураг! Зарим ойлголтуудад түүнийг илэрхийлэх зураг гарч ирэх бөгөөд зургууд бүрэн дүүрэн илэрхийлж чадахгүй, хэсэгчлэн эсвэл тухайн ойлголттой холбоогүй байж болно.

Алгасах! Үг онооход итгэлтэй биш байгаа бол АЛГАСАХ товчийг дарж дараагийн ойлголтыг дуудаж ажиллана уу. Энэ тохиолдолд тэр ойлголт дахиж танд харагдахгүй.

Хэрэглэгчийн интерфэйсийн зохиомж: Синсет орчуулах даалгаврын интерфэйсийн (Зураг 3.20) зохиомжид тавих шаардлагыг дор жагсаав.

1. Интерфэйс нь өөртөө заавраа агуулж баримтжуулсан байх;
2. Товч ба энгийн ойлгомжтой байх;
3. Хэрэглэгч үгзүйн нэгж оноож мэдэхгүй бол үгийг алгасаж дараагийн даалгавар луу шилжих боломжтой байх;
4. Зааварт дурдсан хориглолтуудыг шалгадаг байх бөгөөд шаардлага хангасан тохиолдолд үгийн сангийн нэгжийг хүлээн авдаг байх;



Зураг 3.20 Синсет орчуулах даалгаврын хэрэглэгчийн интерфэйс²³

Дээрх зурагт ойлголтыг илэрхийлэх зургийг Стэнфорд болон Принстоны их сургуулийн хамтарсан судалгааны хамтлагийн хөгжүүлсэн ImageNet²⁴ [70] зургийн онтологиос авч хэрэглэсэн болно. ImageNet нь ВөрдНэтийн синсетэд харгалзах зургийг MTurk платформыг ашиглан олны хүчээр цуглуулж, үнэлүүлж гаргаж авсан, одоогоор 21,841 синсетийн 14.1 сая зургийг хадгалсан зурган өгөгдлийн сан юм. Бодит болон хийсвэр зүйлийг илэрхийлэх ойлголтыг фото болон зурмал зургаар илэрхийлэх боломжтой тул синсетийг орчуулахад түүнийг илэрхийлэх зургийг туслах маягаар ашигласан.

3.3.1.2 Синсет үнэлэх

Синсет орчуулах үед цугларсан тухайн синсетийн үгсээс давхардаагүй үгсээр синсетийн санал болгосон үгсийг (candidate words) бүрдүүлж вэб хэрэглэгчдээр

²³ <http://crowdcrafting.org/project/mongolian-lkc>

²⁴ <http://image-net.org>

үнэлүүлнэ. Ингэж үнэлэхэд вэб хэрэглэгчид *Зөв, Буруу, Мэдэхгүй* гэсэн тэмдэглэгээг хийнэ.

Синсет үнэлэх даалгаврын зохиомжид тавих шаардлага

1. Вэб хэрэглэгчид үгийг үнэлэхдээ англи синсетийг харьцуулж харах шаардлагатай. Энэ нь үнэлэх үгс ямар синсетийг орчуулаад гарч ирсэн үгс болохыг харуулна.
2. Вэб хэрэглэгчид санал болгосон олон үгсийг бүгдэд нь үнэлгээ өгнө. Олон үгс *Зөв* эсвэл *Буруу* эсвэл *Мэдэхгүй* гэсэн тэмдэглэгээ авч болно.
3. Үнэлэх үгсийн дунд дүйцэлгүй ойлголт гэсэн тэмдэглэгээ байх бөгөөд үүнийг ч бас зөв, буруу, мэдэхгүй гэдэг тэмдэглэгээнүүдээс аль нэгийг сонгож тэмдэглэнэ. Хэрэв дүйцэлгүй ойлголтыг зөв гэж тэмдэглэсэн бол бусад бүх үгийг буруу гэж тэмдэглэх ёстой. Хэрэв дүйцэлгүй ойлголтыг буруу гэж тэмдэглэсэн бол санал болгосон үгсээс ядаж нэг нь зөв тэмдэглэгээтэй байх ёстой.

Олны хүчээр гүйцэтгэх ажлын асуулт болон зааврыг дараах байдлаар тодорхойлсон.

Даалгаврын асуулт: Дараах ойлголтод тохирох Монгол орчуулгыг үнэлнэ үү.

Заавар: Орчуулга хийхдээ дараах зааврын дагуу хийнэ.

1. Орчуулсан үг бүрийг ЗӨВ, БУРУУ, МЭДЭХГҮЙ сонголтоос аль нэгийг сонгож үнэлнэ.
2. Зөв бичих үгсийн дүрмийн алдаатай тохиолдолд БУРУУ гэж үнэлнэ.
3. Орчуулсан үгсийн дунд GAP гэсэн үг байвал энэ нь тухайн ойлголтыг орчуулах боломжгүй гэсэн үг юм. Хэрэв та GAP-ийг ЗӨВ гэж үнэлсэн бол бусад үгс нь зөвхөн Буруу эсвэл Мэдэхгүй гэж үнэлэгдэх ёстой.

Дараах нөхцөлд таны үнэлгээг хүлээж авахгүй

1. Санамсаргүйгээр сонгож үнэлбэл
2. Бүх үгийг дан нэг янзаар (дан зөв эсвэл дан буруу) үнэлбэл

Учир нь таны өгсөн үнэлгээг бусад хүмүүсийнхтэй харьцуулж тооцдог тул санамсаргүй болон худал үнэлгээ өгснийг ялгаж чадна.

Хэрэглэгчийн интерфэйсийн зохиомж: Синсет үнэлэх даалгаврын интерфэйсийн (Зураг 3.21) зохиомжид тавих шаардлагыг дор жагсаав.


1. Ямар синсетийг орчуулсан болохыг харуулах;

2. Синсетэд санал болгосон үгс дээр дарж хялбар байдлаар үнэлгээ өгөх;
3. Зааварт дурдсан хориглолтуудыг шалгадаг байх бөгөөд шаардлага хангасан тохиолдолд үнэлгээг хүлээн авдаг байх;

Дараах ойлголтод тохирох Монгол орчуулгыг үнэлнэ үү?

cloister; religious residence

residence that is a place of religious seclusion (such as a monastery)



Санамж: Зураг нь тухайн ойлголтыг ЗӨВ ИЛЭРХИЙЛЭХГҮЙ байж болно!

Үнэлэх үгс:

хийд	БУРУУ
мөргөлийн газар	ЗӨВ
оршин суух газар	БУРУУ
сүм	БУРУУ
уран барилга	БУРУУ
харш	
шашны газар	мэдэхгүй
шашны сууц	БУРУУ
шашны хотхон	ЗӨВ БУРУУ мэдэхгүй

ИЛГЭЭХ

Үнэлэх ажлын дугаар: 972845

Та нийт 1 -с 0 -г үнэлсэн. Заавар

ХЭЛЭЛЦЭХ

Зураг 3.21 Синсет үнэлэх даалгаврын хэрэглэгчийн интерфэйс²⁵

Бид энэ хоёр үет олны хүчний төслийг Crowdcrafting²⁶ платформыг ашиглан зохион байгуулсан. Энэ нь сайн дурын оролцогчдоор олны хүчний төслүүдийг хийлгэж болдог платформ юм. Энэ платформд HTML, CSS, ЖаваСкрипт хэлээр хүний оюуны даалгаврын хэрэглэгчийн интерфэйсийг хэрэгжүүлсэн бөгөөд олны хүчээр гүйцэтгэсэн даалгавараас цугларсан өгөгдлийг CSV, JSON форматаар гаргаж авсан. Улмаар Жава хэлээр өгөгдлийг задалж унших програм бичиж PostgreSQL өгөгдлийн сангийн менежмент системд импортлож оруулсан ба өгөгдлийг EMC-ийн өгөгдөлтэй нэгтгэж өгөгдлийн шинжилгээ хийх, Флайс каппа, Криппендорфийн альфа зэрэг статистик хэмжүүрээр тооцоолоход шаардлагатай өгөгдлийг гаргах зэрэг ажлуудад бэлтгэсэн болно.

²⁵ <http://crowdcrafting.org/project/mongolian-lkc-evaluation>

²⁶ <https://crowdcrafting.org>

3.3.2 Синсетийг үнэлэх ба үгсийг нэгтгэх

Нэг синсетийн үгсийг өөр өөр вэб хэрэглэгчид харилцан адилгүй үнэлгээ өгнө. Эдгээр үнэлгээнүүдээс синсетэд тохирох зөв үгсийг олох шаардлагатай. Үүний тулд 2.3.1-р дэд бүлэгт танилцуулсан Флайс каппа болон Криппендорффийн альфа зэрэг нэрлэн-харьцуулах өгөгдөл дээр олон тооны санал өгөгчтэй байхад ажиллаж чаддаг статистик хэмжүүрийг ашиглан тухайн синсетийн үгсэд үнэлгээ өгсөн вэб хэрэглэгчдийн дотоод санал нийцлийг тооцсон. Хэрэв санал нийцэл өндөр байвал тухайн синсетийг вэб хэрэглэгчид сайн/зөв үнэлсэн. Өөрөөр хэлбэл, синсетэд санал болгосон үгсийг зөв бурууг нь ялгаж чадсан гэсэн үг. Дараа нь сайн нийцтэй үнэлэгдсэн синсетээс дийлэнх олонхын саналаар *Зөв* үнэлгээ авсан үгсийг ялган авах замаар эх синсетэд хамгийн оновчтой тохирох зорилгын синсетийг тодорхойлж үр дүнг нэгтгэнэ.

Дараах бодит синсетийн жишээн дээр Флайс каппа болон Криппендорффийн альфаг бодож үзье.

hovel, hut, hutch, shack, shanty -- *small crude shelter used as a dwelling*

3.3.2.1 Флайс каппа

Дээрх англи синсетийг орчуулгад 5 оролцогчид давхардаагүй 4 үгсийг санал болгосон (Хүснэгт 3.6).

Хүснэгт 3.6 Синсетийн орчуулгад санал болгосон үгс

үгс	санал болгосон оролцогчдын тоо
овоохой	4
оромж	2
урц	3
саравч	3

Үүнээс 4 оролцогч *овоохой*, 2 нь *оромж* гэдэг үгийг санал болгосон байна. Тэмдэглэж хэлэхэд хоёр өөр оролцогч тус тусдаа ажиллах ба нэг үгийг давхардуулан орчуулах боломжтой. Дараа нь оролцогчдоос санал болгосон үгсийг *Зөв*, *Буруу*, *Мэдэхгүй* гэж 3 ангилуулахад Хүснэгт 3.7-д үзүүлсэн саналыг өгчээ. Энд 5 санал өгөгч (n) 4 үгсийг (N) 3 ангилалд (k) оруулах санал өгсөн байна. Ангиллыг баганад, үгсийг мөрөнд үзүүлэв.

Хүснэгтийн нэг нүдэнд *i*-р үгийг *j*-р ангилалд оруулах санал өгсөн нийт санал өгөгчийн тоо байна. Эндээс p_1 -ийг (3.2)-д үзүүлснээр бодно.

$$p_1 = \frac{4 + 5 + 3 + 0}{20} = 0.600 \quad (3.2)$$

Хүснэгт 3.7 Синсетийн орчуулгад санал өгсөн байдал

		ангилал (k)				
		n_{ij}	Зөв	Буруу	Мэдэхгүй	P_i
үгс (N)	1	овоохой	4	1	0	0.600
	2	оромж	5	0	0	1.000
	3	урц	3	2	0	0.400
	4	саравч	0	3	2	0.400
		Нийт санал	12	6	2	$Nn = 20$
		p_j	0.600	0.300	0.100	2.400

P_1 -ийг (3.3)-ийн дагуу тооцсон шиг бүх үгсийн хувьд бодно.

$$P_1 = \frac{1}{5(5-1)}(4^2 + 1^2 + 0^2 - 5) = 0.600 \quad (3.3)$$

Одоо \bar{P} -ийг (3.4)-д үзүүлснээр, \bar{P}_e -г (3.5)-д үзүүлснээр тус тус олно.

$$\bar{P} = \frac{1}{4}(0.600 + 1.000 + 0.400 + 0.400) = 0.600 \quad (3.4)$$

$$\bar{P}_e = 0.600^2 + 0.300^2 + 0.100^2 = 0.460 \quad (3.5)$$

Ингээд Флайс каппаг бодвол дараах шиг үр дүн гарна.

$$k = \frac{0.600 - 0.460}{1 - 0.460} = 0.259 \quad (3.6)$$

Каппагийн утга 0.259 бөгөөд 0.21 – 0.40 хооронд байх тул мэдэгдэхүйц санал нийцэлтэй гэж үзнэ.

Хэрэв санал өгөгчид бүх үгсийг *Зөв* эсвэл *Буруу* эсвэл *Мэдэхгүй* гэж зөвхөн нэг ангилсан тохиолдолд \bar{P} болон \bar{P}_e утгууд 1 болно. Иймд (2.7) томъёоны хүртвэр $\bar{P} - \bar{P}_e = 0$ болж каппа бодогдохгүй. Харин синсетийг үгсийг ангилах үед орчуулахад хялбар синсет 1 эсвэл 2 үгтэй байж болох бөгөөд бүх санал өгөгчид синсетийн үгсийг бүгдийг нь зөв эсвэл буруу гэж ангилах магадлалтай. Энэ тохиолдолд санал өгөгчид 100% санал нийлсэн гэж үзээд каппагийн утгыг 1.0 гэж үзэж болно.

3.3.2.2 Кrippendorffийн альфа

Хүснэгт 3.8-д дээрх синсетийн 4 үгсэд 5 санал өгөгчийн саналын үнэн өгөгдлийн матрицыг үзүүлэв. Энд санал нь 1 бол *Зөв*, 0 бол *Буруу*, 2 бол *Мэдэхгүй* гэсэн ангилал болно. Энд бүх санал өгөгчид санал өгсөн байна. m_u бол u үгэнд санал өгсөн хүний

тоо. Хүснэгт 3.9-д синсетийн үгсийн ангиллын тохиролцлын хүснэгтийг үзүүлэв. Энд *овоохой* үгийн хувьд нийт $c-k$ ангиллын $5(5-1)=20$ хос байна. Үүнээс 1-1 хос $4(4-1)=12$, 1-0 хос 4, 0-1 хос 4 байна. *урц* үгийн ангиллын нийт 20 хосоос 1-0 нь 6, 0-1 нь 6 байна. Бусад үгсийн хувьд 1-0 болон 0-1 хосын тохиолдол байхгүй.

Хүснэгт 3.8 Синсетэд санал өгсөн үнэн өгөгдлийн матриц

Санал өгөгч/Үгс	овоохой	оромж	урц	саравч
Санал өгөгч 1	1	1	1	0
Санал өгөгч 2	1	1	1	0
Санал өгөгч 3	1	1	1	0
Санал өгөгч 4	1	1	0	2
Санал өгөгч 5	0	1	0	2
m_{ui}	5	5	5	5

Иймд нийт 1-0 хосын тоо 10 байна. Харин 0-0 хосын хувьд *овоохой* үгийн хувьд 0, *урц* үгийн хувьд 2, *саравч* үгийн хувьд 6 нийт 8 байна. Тиймээс $n_0=24$, $n_1=48$, $n_2=8$ бөгөөд $n=80$ болно.

Хүснэгт 3.9 Синсетийн үгсийн ангиллын тохиролцлын хүснэгт

	0	1	2	
0	8	10	6	24
1	10	38	0	48
2	6	0	2	8
n	24	48	8	80

Эндээс альфаг (3.7) томъёогоор бодно. Энд зөрүүгийн функцийг тооцоогүй болно.

$$\alpha = 1 - \frac{D_o}{D_e} = \frac{A_o - A_e}{1 - A_e} = \frac{(n-1)\sum_c o_{cc} - \sum_c n_c(n_c-1)}{n(n-1) - \sum_c n_c(n_c-1)} \quad (3.7)$$

$$\alpha = \frac{(80-1)(8+38+2) - [24(24-1) + 48(48-1) + 8(8-1)]}{80(80-1) - [24(24-1) + 48(48-1) + 8(8-1)]} = 0.269 \quad (3.8)$$

Синсетийн үгсийг сонгож авахын тулд *Зөв*, *Буруу* гэдэг ангилал голлох үүрэгтэй тул *Мэдэхгүй* гэсэн ангиллыг ангилал биш, орхисон өгөгдөл гэж үзэж болох юм. Үнэн хэрэгтээ энэ ангиллыг синсетийн үгийг сонгож авахад ашиглаагүй. Иймд бид *Зөв*, *Буруу* гэсэн хоёрхон ангилалтай гэж үзэж болно. Олны хүчээр гүйцэтгүүлж байгаа ажлын хувьд оролцогчид *Мэдэхгүй* гэж тэмдэглэх эсвэл алгасах гэх маягийн сонголтуудыг олгох нь зүйтэй байдаг. Учир нь үнэхээр гүйцэтгэж чадахгүй ажилд бид буруу өгөгдөл авахаас сэргийлж буй хэрэг. Энэ тохиолдолд синсетийн үгсийн

ангиллын үнэн өгөгдлийн матрицыг Хүснэгт 3.10-т, тохиролцлын хүснэгтийг Хүснэгт 3.11-д үзүүлэв. Энд 2 гэдэг ангиллыг санал өгөгчид өгөөгүй, өөрөөр хэлбэл, тухайн үгийг ямар нэг ангилалд оруулаагүй ($N \setminus A$) гэсэн утгатай.

Хүснэгт 3.10 Синсетэд санал өгсөн үнэн өгөгдлийн матриц (орхисон өгөгдөлтэй)

Санал өгөгч/Үгс	овоохой	оромж	урц	саравч
Санал өгөгч 1	1	1	1	0
Санал өгөгч 2	1	1	1	0
Санал өгөгч 3	1	1	1	0
Санал өгөгч 4	1	1	0	.
Санал өгөгч 5	0	1	0	.
m_u	5	5	5	3

(2.15) томъёоны дагуу орхисон өгөгдөлтэй үед o_{ck} -г (3.9)-д үзүүлснээр бодно.

$$o_{00} = \frac{0}{5-1} + \frac{0}{5-1} + \frac{2}{5-1} + \frac{6}{3-1} = 3.5 \quad (3.9)$$

Энд *овоохой* болон *оромж* үгийн хувьд саналын 0-0 гэсэн хос байхгүй, *урц* үгийн хувьд 0-0 хос 2, *саравч* үгийн хувьд 6 байгаад 3 хүн санал өгсөн байна. Иймд дараах тохиролцлын хүснэгт гарна.

Хүснэгт 3.11 Синсетийн үгсийн ангиллын тохиролцлын хүснэгт (орхисон өгөгдөлтэй)

	0	1	2	
0	3.5	2.5	.	6
1	2.5	9.5	.	12
2
n	6	12	.	18

Энэ тохиролцлын хүснэгтээс альфаг (3.10)-д үзүүлсэн шиг бодно.

$$\alpha = \frac{(18-1)(3.5+9.5) - [6(6-1) + 12(12-1)]}{18(18-1) - [6(6-1) + 12(12-1)]} = 0.410 \quad (3.10)$$

Мэдэхгүй гэсэн ангиллыг ангилал биш гэж үзсэн тохиолдолд альфагийн утга өсөж байна. Энд ангиллын тоо буурч 2 болж санал өгөгчдийн санал хуваагдах магадлал буурч байгаа учир альфа нэмэгдэнэ.

3.3.3 Туришилт, үр дүн

Туршилтаар Орон зайн айн англи хэлээр илэрхийлсэн 947 синсетийг нийт 77 вэб хэрэглэгч орчуулж, 75 вэб хэрэглэгч үнэлгээ өгсөн (Хүснэгт 3.12). Нэг синсетийг 5 өөр

вэб хэрэглэгчээр орчуулуулж 5 өөр вэб хэрэглэгчээр үнэлүүлсэн. $949 * 5 = 4,745$ ХОД-ыг үе шат тус бүрд гүйцэтгүүлсэн болно. К.Каллисон-Бөрч [56] санал өгөгчийн тоо 5 байхад хамгийн өндөр BLEU оноотой орчуулгыг гаргаж авсан тул нэг даалгавар дээр ажиллах вэб хэрэглэгчийн тоог 5-аар авсан юм. Эхний үед вэб хэрэглэгчид нийт 8,339 үгийг синсенийн орчуулгад санал болгосон. Үүнээс давхардаагүй нь 4,854 байв. Нэг үгийг нэг синсетэд эсвэл өөр синсетэд олон хэрэглэгч санал болгож болно. Жишээ нь, *агуй* гэдэг үгийг 4 синсетэд санал болгосон байна. Эцэст нь 947 синсенийн 6442 үгсийг олны хүчээр орчуулуулж авсан.

Хүснэгт 3.12 Хүний оюуны даалгаврын гүйцэтгэлийн тоон үзүүлэлт

№	Үзүүлэлт	Утга
1	Даалгаварт өгсөн синсенийн тоо	947
2	Орчуулгын даалгаврын оролцогчид	77
3	Үнэлгээний даалгаварын оролцогчид	75
4	Орчуулгын ХОД	4,745
5	Үнэлгээний ХОД	4,745
6	Орчуулгын ХОД-аас цугларсан нийт үгс	8,339
7	Орчуулгын ХОД-аас цугларсан давхардаагүй үгс	4,854
8	Үнэлгээний ХОД-т өгсөн саналын нийт тоо	45,117

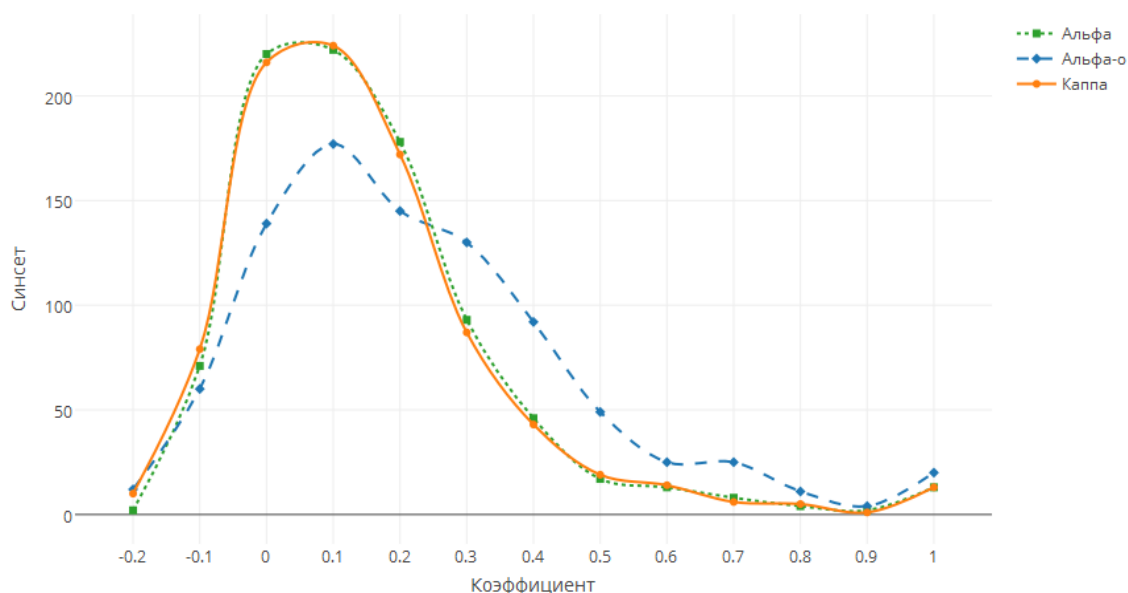
Олны хүчээр гүйцэтгэсэн орчуулгын чанарыг үнэлэхдээ жишиг үгийн сантай харьцуулсан. Жишиг үгийн санг бэлдэхдээ мэргэжилтнүүдээр онтологи нутагшуулах аргачлалын үр дүнд бий болсон 943 синсенийн 1,436 үгсийг ашигласан. Мөн олны хүчээр гүйцэтгүүлсэн үгсийг 3 мэргэжилтнээр *Зөв*, *Буруу* гэж үнэлүүлэн үгсийн сан гаргаж авсан. Энэ санд 889 синсенийн 2,461 үгсээс 1,813 *Зөв* үнэлгээтэй, 647 нь *Буруу* үнэлгээтэй байв. Үлдсэн 58 (947-889) синсетэд дийлэнх олонхын санал авсан ямар нэг үг байгаагүй болно. Энэ хоёр санг нэгтгэж жишиг үгийн санг гаргаж авсан болно (Хүснэгт 3.13).

Хүснэгт 3.13 Жишиг сангийн хэмжээ

№	Олны хүчний төрөл	Синсет	Үгсийн тоо
1	Мэргэжилтэн	943	1,436
2	Вэб хэрэглэгчид	889	1,813
	Жишиг сан	943	2,627

Эндээс вэб хэрэглэгчдээр гүйцэтгүүлсэн синсенийн орчуулгад санал нийлсэн байдлыг Зураг 3.22-д үзүүлэв. Альфа болон каппагийн санал нийцэл нь 0.00 – 0.09 хооронд (хэвтээ тэнхлэгийн 0 утга дээр тэмдэглэсэн) байх синсенийн тоо 220, орхисон

өгөгдөлтэй альфагийн хувьд 139 байна. Альфаг орхисон өгөгдөлтэй тооцох үед санал нийцэлтэй синсетийн тоо өсч байна. Өөрөөр хэлбэл, энэ аргаар хэмжихэд санал нийцэлтэй гарсан синсетийн тоо өсөх ба олон синсетээс олон 3өв үнэлгээтэй үгсийг сонгож авах боломж ихсэнэ.

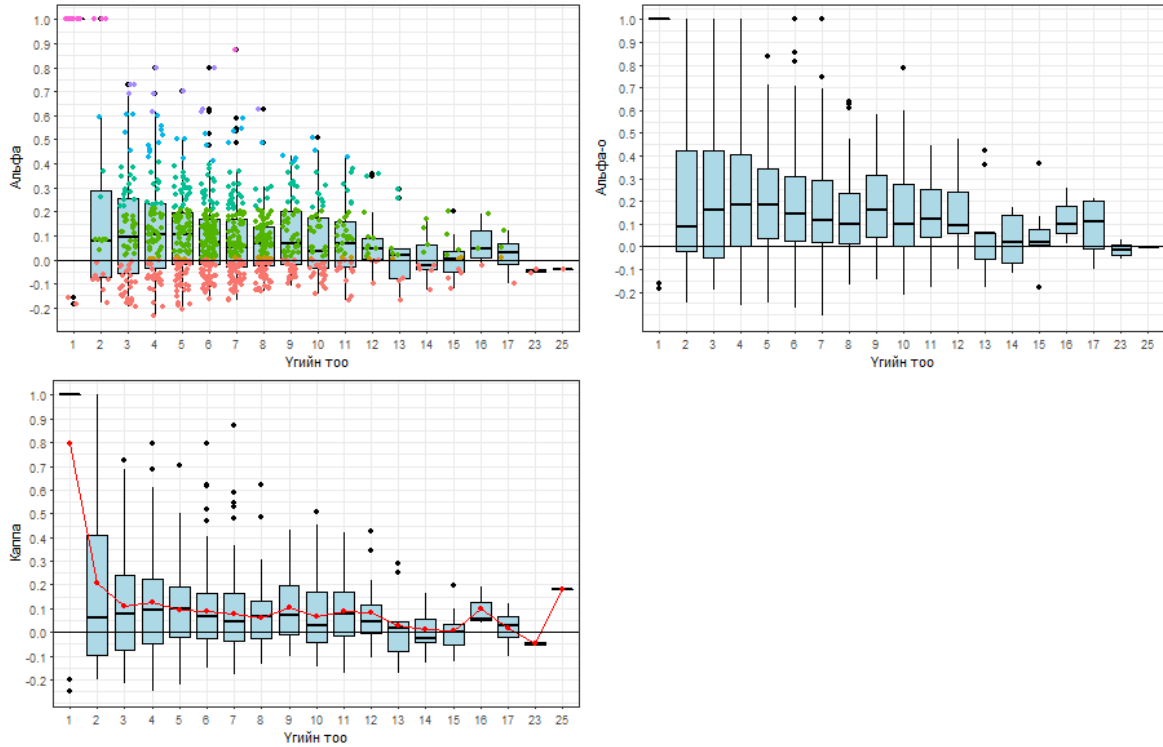


Зураг 3.22 Санал нийцлийн хэмжүүрийн утгаас хамаарсан синсетийн давтамж

Төгс санал нийцэлтэй синсет 20 орчим буюу нийт синсетийн 2.1 хувь байсан бол нийт синсетийн 21.1 хувь нь санал нийцэлгүй гарсан.

Зураг 3.23-оос харахад синсетийн үгийн тоо өсөхөд санал нийцлийн коэффициентын утга өсөх эсвэл буурах хандлага ажиглагдахгүй байна. Өөрөөр хэлбэл, олон үгтэй синсетийн санал нийцэл цөөн үгтэй синсетийнхтэй ижил байна. Үүнийг синсетийн үгийн тооноос хамаарсан каппагийн графикт дундаж утгуудыг дайрсан шулуунаас харж болох ба бусад хоёр аргын хувьд ижил дүр зураг харагдаж байна.

Олны хүчээр гүйцэтгэсэн синсет бүрийн санал нийцлийг альфа, альфа-о, каппа зэрэг 3 статистик хэмжүүрээр хэмжиж дийлэнх олонхын санал авсан үгсийг жишиг сантай харьцуулж үзсэн. Улмаар синсетийн тохирлын үзүүлэлтүүдийг тооцсон болно (Зураг 3.24).



Зураг 3.23 Синсетийн үгсийн тооноос хамаарсан санал нийцэл

Тохирлын үзүүлэлтэд тохирол (precision), тусгал (recall), ф1-оноо (f1-score) зэрэг багтана.

$$precision = \frac{tp}{tp + fp} \quad (3.11)$$

$$recall = \frac{tp}{tp + fn} \quad (3.12)$$

$$f\text{-measure} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (3.13)$$

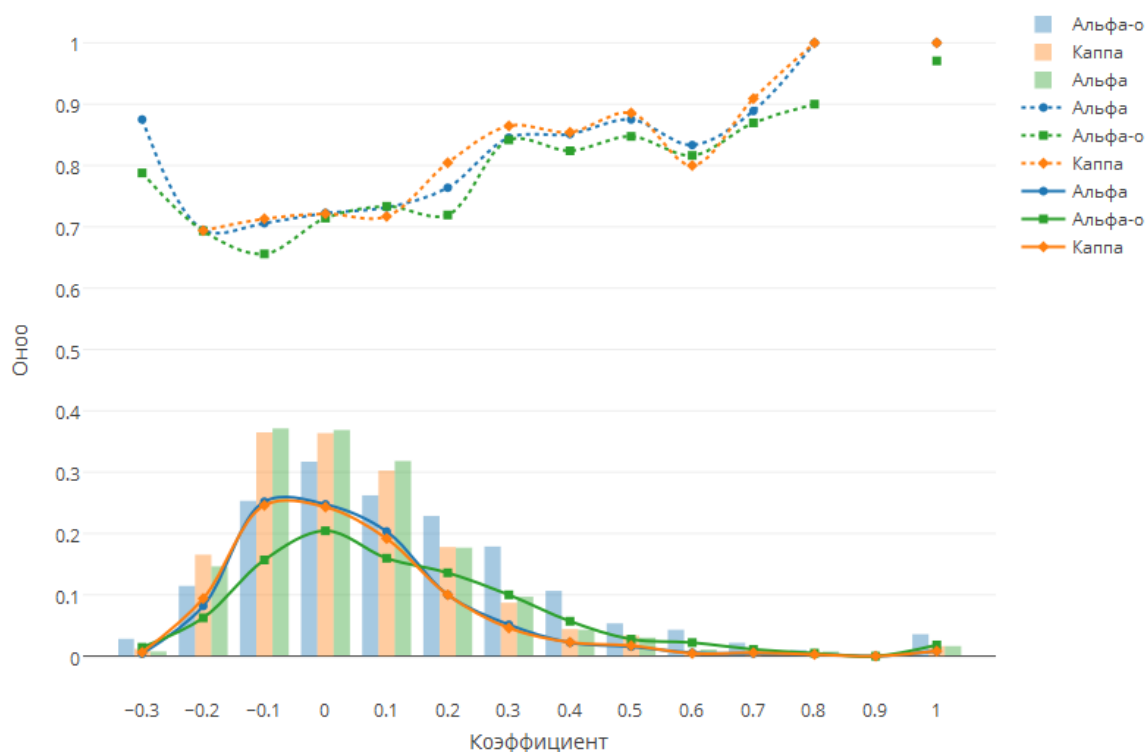
Энд tp нь жишиг санд зөв олдсон үгсийн тоо, fp буруу үгсийг зөв гэж үзсэн тоо юм. Иймд тохирол нь олны хүчээр зөв орчуулсан үгсийг нийт зөв орчуулсан үгс болон олж чадаагүй зөв үгсийн нийлбэрт харьцуулсан харьцаа юм. Өөрөөр хэлбэл, олны хүчээр орчуулсан нийт үгсээс зөв гэж сонгосон үгсийн хэд нь жишиг санд зөв тохирч байгааг илэрхийлнэ. fn гэдэг нь олж чадаагүй зөв үгсийн тоо болно. Тусгал нь жишиг санд байгаа нийт зөв үгсийн хэд нь зөв олдсон гэсэн харьцаа юм. Харин ф1-оноо нь тохирол болон буцаалтын харьцаагаар хэмжигддэг хэмжигдэхүүн юм.

Энэ туршилтыг дараах нөхцөлд хийж гүйцэтгэсэн болно.

- Эх хэл Англи, зорилтот хэл Монгол байна.

- Нэг синсетэд олон үг оноох боломжтой.
- Нэг ХОД-ыг нэг оролцогч гүйцэтгэнэ.
- Оролцогчид Англи хэлний A2 түвшний мэдлэгтэй байна.
- Оролцогчид унаган монгол хэлтэн байна.
- Оролцогчид 17-оос дээш настай байна.

Зураг 3.24-т синсетийн санал нийцлийг альфа, альфа-о болон каппагийн аргаар бодож санал нийцлийн тухайн утгад олдсон синсетүүдээс дийлэнх олонхын саналаар сонгож авсан үгсийг жишиг сантай харьцуулахад тохирол, тусгал, ф1-оноо ямар байгааг харуулсан болно. Хэвтээ тэнхлэгийн дагуу санал нийцлийн коэффициентын утгууд байгаа бөгөөд 0.09 алхамтай. Эдгээр утгууд дээрх багануудаар ф1-оноог, дугуй цэгээр тохирол, гурвалжин цэгээр тусгалыг үзүүлсэн. Үргэлжилсэн шулуунаар тусгалыг, тасархай шулуунаар тохирлыг илэрхийлсэн. Тухайн утга дээр байгаа 3 баганын зүүн талынх нь альфа-о, голын багана каппа, баруун захынх нь альфагийн ф1-оноог илэрхийлнэ. Жишээ нь, 0.1-0.19 хооронд санал нийцлийн зэрэгтэй синсетүүдээс дийлэнх олонхын саналаар сонгож авсан үгсийн ф1-оноо альфагийн хувьд 0.32, альфа-о 0.26, каппа 0.30 байна.



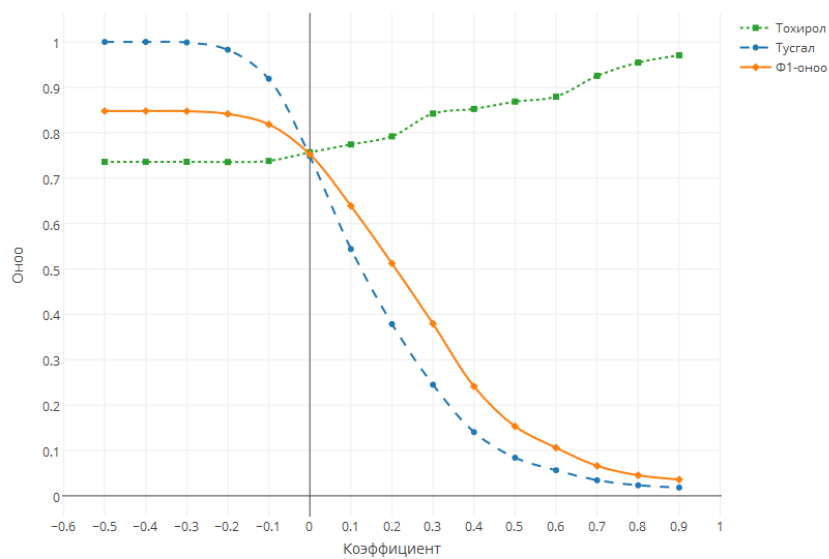
Зураг 3.24 Санал нийцлийн коэффициентын утгууд дээрх синсетийн тохирлын үзүүлэлтүүд

Тохирлын хувьд, санал нийцлийн утга 0-0.1 байх синсетүүдийн дийлэнх олонхийн саналаар сонгосон үгсийг жишиг сантай харьцуулж үзэхэд альфагийн хувьд 0.72, альфа-о-ийн хувьд 0.71, каппагийн хувьд 0.72 байна. Зураг 3.24-өөс харахад санал нийцлийн коэффициентийн утга өсөхөд тохирлын үзүүлэлтүүд өсөх хандлагатай байна. Энэ нь синсетийн үгсийг үнэлсэн оролцогчдын дотоод санал нийцэл өндөр байхад дийлэнх олонхийн саналаар сонгож авах үгс нь жишиг санд зөв олдож буйг илэрхийлж байна. Мөн синсетийн санал нийцлийн зэрэг өсөх тусам тохирол өсөж, тусгал буурна. Өөрөөр хэлбэл, синсетийн үгсийг зөв, буруу, мэдэхгүй гэж санал өгсөн вэб хэрэглэгчдийн дотоод санал нийцэл өндөр байх тусам зөв орчуулгыг гарган авч чадна гэдгийг баталж байна.

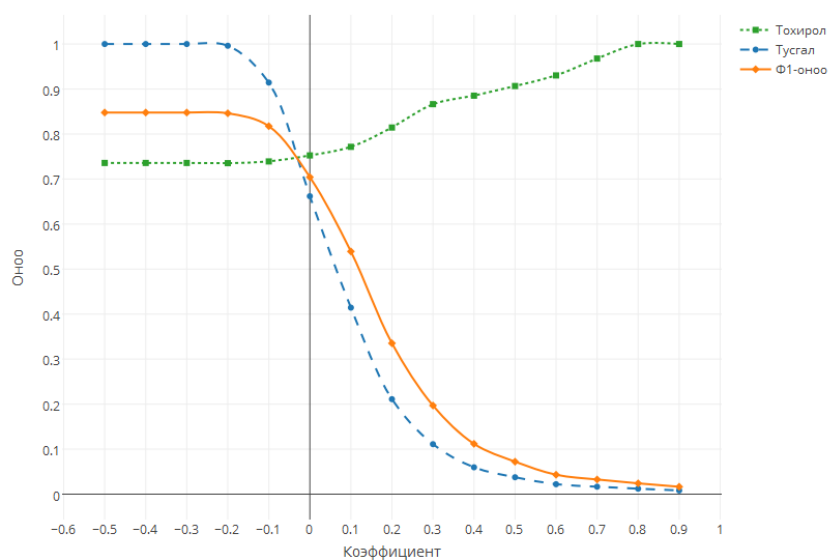
Зураг 3.25-д санал нийцлийн коэффициент тодорхой утгаас их байх синсетүүдийг сонгон авч тус бүрт нь тооцсон тохирлын үзүүлэлтүүдийг харуулсан. Ф1-онооны хувьд хамгийн өндөр нь альфагаар бодсон санал нийцлийн арга байсан бол үлдсэн хоёр аргын хувьд тохирлын үзүүлэлтүүд хоорондоо бараг ижил байна. Гэвч тохирлын хувьд энэ хоёр арга альфагаас илүү өндөр, тусгалын хувьд бага байгаа юм. Хэрэв онтологи нутагшуулалтын ажлын зорилгод аль болох олон зөв үгсийг гаргаж авах нь чухал гэж үзвэл өндөр тохиролтой аргууд болох альфа-о болон каппаг ашиглах нь зүйтэй харагдаж байна.

Альфа аргын хувьд (Зураг 3.25.а) тохирол тусгал хоёрын огтлолцлын цэг коэффициентын 0 утга дээр байгаа. Харин альфа-о (Зураг 3.25.б) болон каппагийн (Зураг 3.25.в) хувьд үл ялиг бага буюу -0.06 орчимд байна. Ийм тохиолдолд тохирол болон тусгалын харьцаа хамгийн оновчтой юм. Харин тохирлын хувьд -0.1-ээс дээш санал нийцэлтэй байхад өсөж эхэлж байгаа тул санал нийцлийн босго утгыг энэ утгаар тогтоож болно.

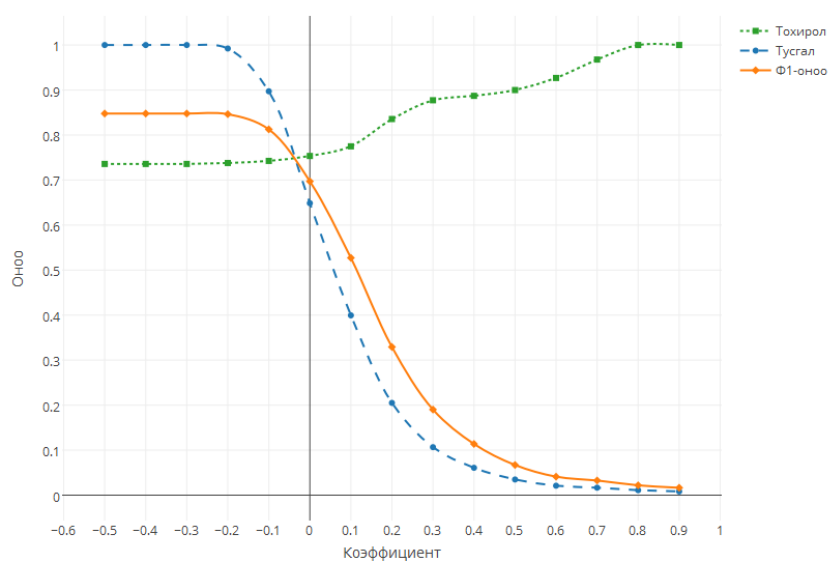
Хэдийгээр онолын хувьд санал нийцлийн коэффициентын утга 0-ээс их үед санал нийлсэн гэж үздэг боловч бидний туршилтад үүнээс бага байхад тухайн синсетийг үнэлсэн вэб хэрэглэгчид синсетийн үгсийн зөв бурууг зөв ялгаж чадсан гэж ойлгож болно. Учир нь коэффициентын утга санал өгөх зүйлсийн тоо, ангиллын тоо ихсэх тусам буурч байдаг.



а. Кrippendorffийн альфа



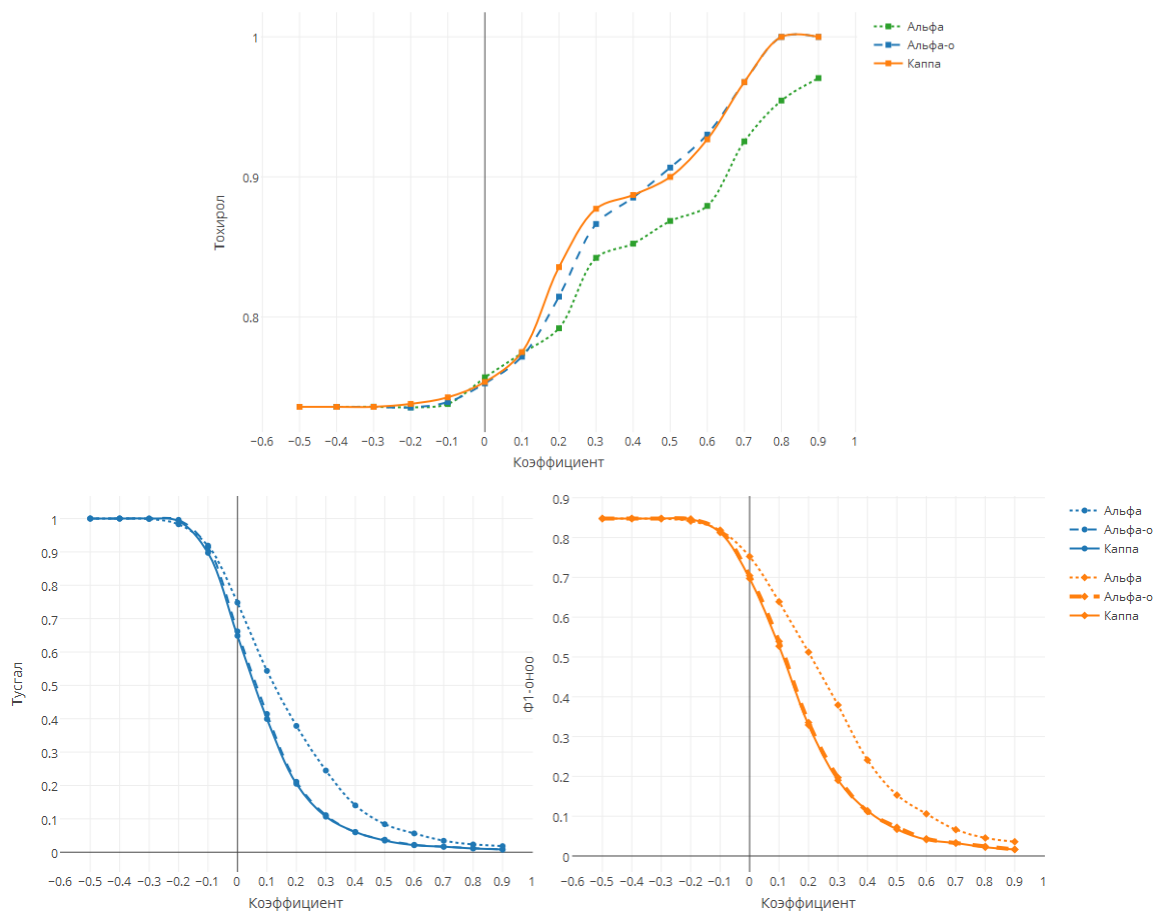
б. Кrippendorffийн альфа (орхисон өгөгдөлтэй үед)



в. Флайс каппа

Зураг 3.25 Санал нийцлийн хэмжүүрүүдийн тохирлын үзүүлэлтүүд

Хэлний цахим нөөц багатай хэлний хувьд эхний зорилго бол аль болох богино хугацаанд олон синсетийг нутагшуулах нь чухал учраас энэ хоёр аргыг сонгох зүйтэй байна.



Зураг 3.26 Тохирлын үзүүлэлтүүдийн хоорондын харьцуулалт

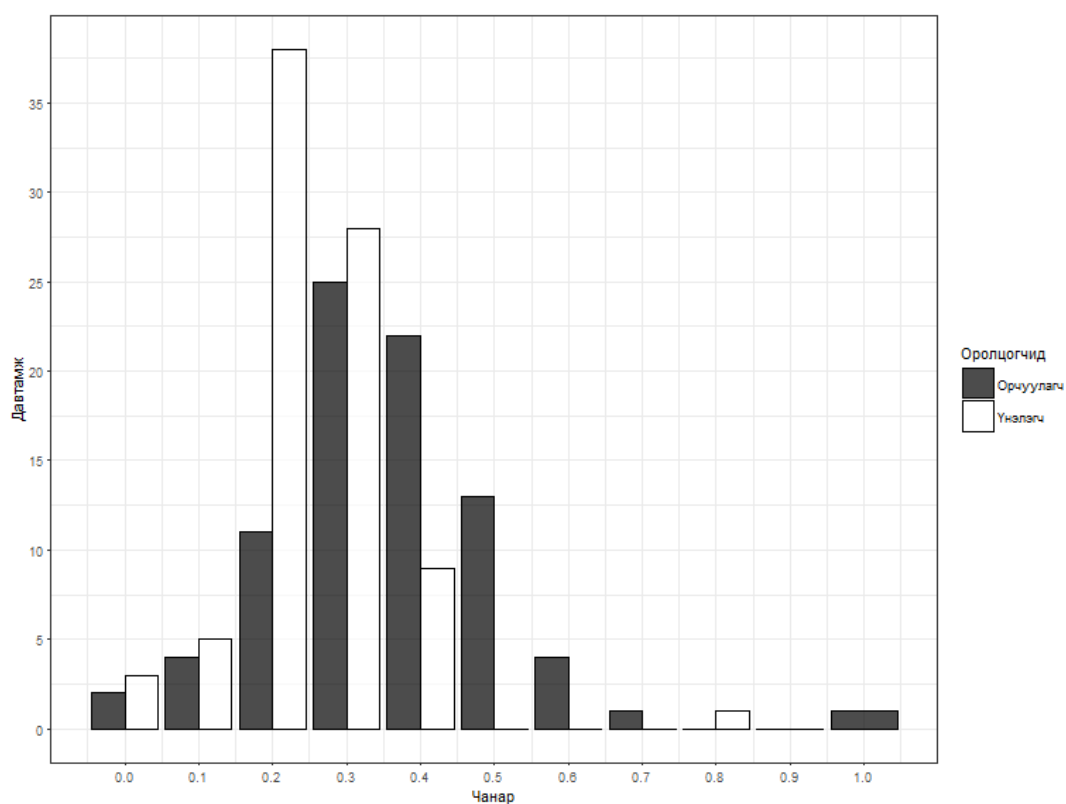
Хэрэв тусгалыг өсгөж f_1 -оноог сайжруулахын тулд бага хүчээр илүү оновчтой орчуулгыг гарган авахад чиглэсэн арга техникийг оруулж өгөх шаардлагатай. Жишээ нь, синсетийн орчуулгын үед давхардуулан санал болгож буй үгс, синсетийг үнэлэх үед дотоод санал нийцлийг даруй тооцож ажлын гүйцэтгэл сайн эсвэл хангалтгүй оролцогчдыг ялгаж улмаар сайн оролцогчдоор тэдний түвшинд тохирсон синсетийг орчуулуулах ба үнэлүүлэх зэрэг байж болно.

Хүснэгт 3.14-т Зураг 3.25 болон Зураг 3.26-д үзүүлсэн тохирлын үзүүлэлтүүдийн утгыг харуулав. Энэ хүснэгтэд тодруулсан утгуудаас эхлэн тохирлын утгууд тогтмол өсч байгаа. Жишээ нь, альфа аргын хувьд тохирлын утга 0.738, альфа-о-ийн хувьд 0.739, каппагийн хувьд 0.743 байна. Энэ нь санал нийцлийн коэффициентийн утга нь -0.1-ээс байх бүх синсетүүдийг сонгон авч жишиг сантай харьцуулан тооцсон тохирлын утгууд болно.

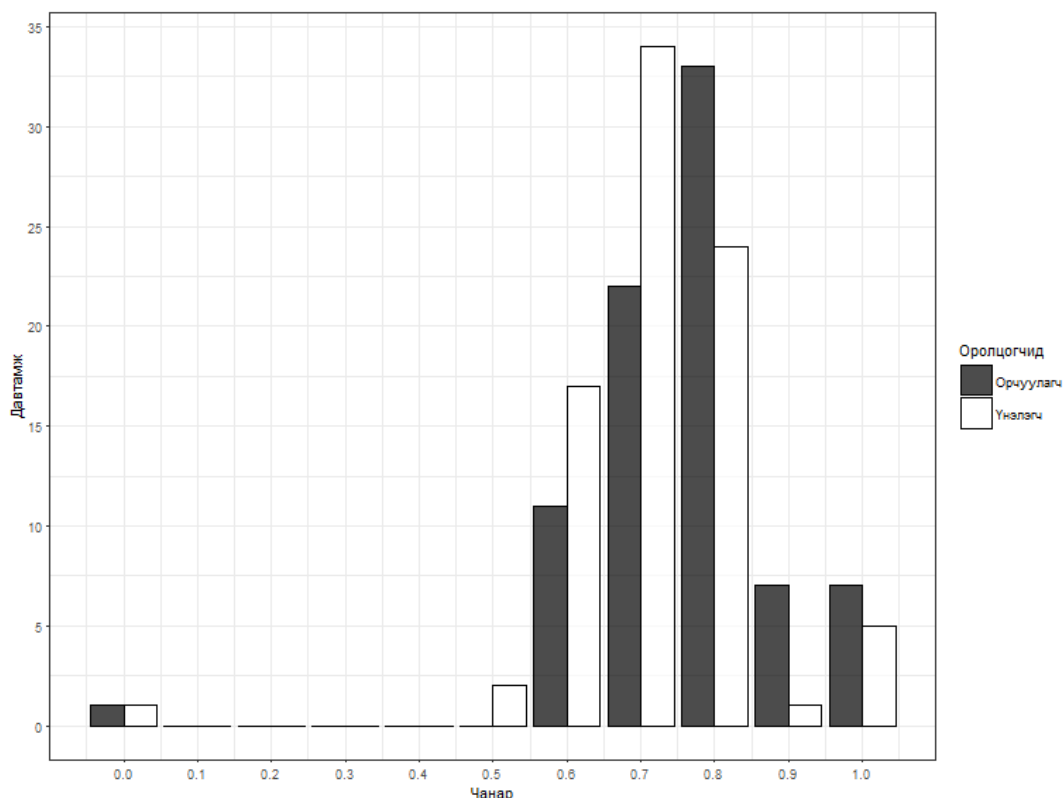
Хүснэгт 3.14 Коэффициентын утга дахь тохирлын үзүүлэлтүүд

№	Коэффициент	Альфа			Альфа-о			Каппа		
		Тохирол	Тусгал	Ф1-оноо	Тохирол	Тусгал	Ф1-оноо	Тохирол	Тусгал	Ф1-оноо
1	-0.5	0.736	1.000	0.848	0.736	1.000	0.848	0.736	1.000	0.848
2	-0.4	0.736	1.000	0.848	0.736	1.000	0.848	0.736	1.000	0.848
3	-0.3	0.736	0.999	0.847	0.736	1.000	0.848	0.736	1.000	0.848
4	-0.2	0.735	0.983	0.841	0.735	0.996	0.846	0.738	0.992	0.846
5	-0.1	0.738	0.919	0.818	0.739	0.914	0.817	0.743	0.897	0.813
6	0	0.757	0.748	0.752	0.753	0.662	0.704	0.754	0.649	0.697
7	0.1	0.774	0.544	0.639	0.772	0.414	0.539	0.775	0.399	0.527
8	0.2	0.792	0.378	0.512	0.814	0.211	0.335	0.836	0.205	0.329
9	0.3	0.842	0.245	0.379	0.866	0.111	0.197	0.877	0.107	0.190
10	0.4	0.852	0.140	0.241	0.885	0.060	0.112	0.887	0.061	0.114
11	0.5	0.869	0.084	0.153	0.907	0.038	0.072	0.900	0.035	0.067
12	0.6	0.879	0.056	0.106	0.930	0.022	0.043	0.927	0.021	0.041
13	0.7	0.925	0.034	0.066	0.968	0.017	0.033	0.968	0.017	0.033
14	0.8	0.955	0.023	0.045	1.000	0.012	0.024	1.000	0.011	0.022
15	0.9	0.971	0.018	0.036	1.000	0.008	0.016	1.000	0.008	0.016

Иймд коэффициентийн утга -0.1-ээс их байхад каппа хэмжүүрээр 0.74 хувийн зөв орчуулгыг гарган авч чадна.



Зураг 3.27 Вэб хэрэглэгчдийн гүйцэтгэсэн ажлын чанар



Зураг 3.28 Дийлэнх олонхын саналаар сонгосон үгсийг орчуулсан болон үнэлсэн вэб хэрэглэгчдийн ажлын чанар

Хэрэв өндөр чанартай орчуулгыг гарган авъя гэвэл коэффициентийн утгыг өсгөж цөөн синсетүүдийг сонгох авах нь зүйтэй. Гэвч энэ тохиолдолд олны хүчээр гүйцэтгүүлсэн ихэнх ажлын үр дүнг авч ашиглахгүй болно. Харин сонгож авах синсетийн тоог ихэсгэхэд жишиг санд зөв олодох үгсийн тоо өсөх ба үүнийг дагаад олж чадаагүй зөв үгсийн тоо ч мөн адил өснө. Ингэснээр тохирол өсөхөд тусгал эсрэгээрээ буурдаг. Үр дүнд нь олон зөв, буруу үгсийг гаргаж авна гэсэн үг. Эцсийн үр дүнг сайжруулахын тулд мэргэжлийн олны хүчээр онтологи нутагшуулах аргачлалын дагуу буруу үгсийг ялгаж болно.

Вэб хэрэглэгчдийн ажлын чанарыг хоёр янзын аргаар үнэлж үзсэн. Зураг 3.27-д синсет орчуулсан вэб хэрэглэгч (translator), синсет үнэлсэн вэб хэрэглэгч (validator) нарын зөв орчуулсан эсвэл үнэлсэн болон буруу орчуулсан эсвэл үнэлсэн ажлын харьцаагаар чанарыг тооцсон. Тухайн вэб хэрэглэгчийн орчуулсан эсвэл үнэлсэн үгс жишиг санд олдсон бол зөв орчуулсан эсвэл үнэлсэн гэж үзнэ. Нийт орчуулсан эсвэл үнэлсэн үгсийн хэдэн хувь нь жишиг санд олдсоныг харуулав. Жишээ нь, синсетийг орчуулах туршилтад оролцсон 77 вэб хэрэглэгчийн 25 нь 30%-ийн чанартай орчуулжээ.

Синсет үнэлэх даалгаварт оролцсон 75 вэб хэрэглэгчийн 28 нь 30%-ийн чанартай үнэлсэн байна. Зураг 3.28-д дийлэнх олонхын саналаар сонгосон синсетийн үгсийг

орчуулсан болон үнэлсэн вэб хэрэглэгчдийн орчуулсан болон үнэлсэн үгсийг жишиг сангийн үгстэй харьцуулж гаргасан ажлын чанарыг үзүүлэв. Жишээ нь, 80%-ийн чанартай орчуулсан 33 орчуулагч, 24 үнэлэгч байна. Ийм маягаар сайн оролцогчид болон муу эсвэл хууран мэхлэгч оролцогчдыг ялгах боломжтой.

Дээрх туршилтуудын хэрэгжүүлэлтийг Хавсралт Д. Тохирлын үзүүлэлтүүдийг тооцоолох R скриптээс үзнэ үү.

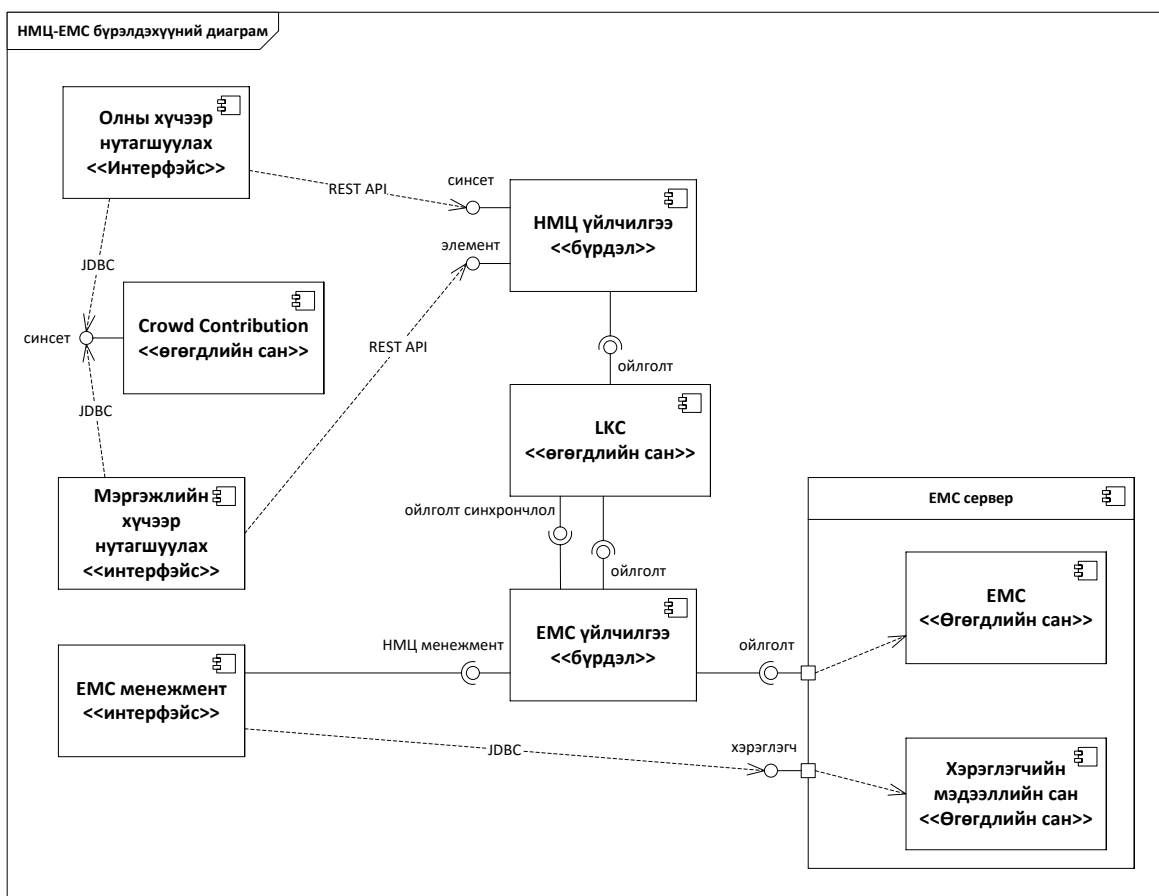
Вэб хэрэглэгчид дийлэнх олонхын саналаар нийт 5 дүйцэлгүй ойлголтыг санал болгосон. Түүнээс 2 нь зөв дүйцэлгүй ойлголт байсан. Эндээс харахад вэб хэрэглэгчид дүйцэлгүй ойлголтыг тодорхойлж чадахгүй байна. Нөгөө талаас дүйцэлгүй ойлголтыг ХОД-ын зохиомжид тодорхой сайн тусгаагүй эсвэл бага үнэлэмжтэй оролцогчид дүйцэлгүй ойлголтод ямар нэг байдлаар синсетийн үгс санал болгосон байж болох юм.

3.4 Үр дүн

Бид мэргэжилтнүүд болон энгийн вэб хэрэглэгчдээр онтологи нутагшуулах аргачлалуудыг хослуулсан аргачлалыг боловсруулсан. Энэ хосолмол аргачлалаар синсетийн үгсийг орчуулах ажлыг нь энгийн вэб хэрэглэгчдээр бусад ажлуудыг нь мэргэжилтнүүдээр гүйцэтгүүлэх юм. Зураг 3.29-д хосолмол аргачлалыг хэрэгжүүлэх системийн бүрэлдэхүүний диаграмыг үзүүлэв. Зураг 3.30-д хосолмол аргачлалын ерөнхий алгоритмыг үзүүлэв. Энд вэб хэрэглэгчдээр орчуулуулах англи синсетийг англи НМЦ-өөс авч олон тооны (туршилтаар 5 өөр) вэб хэрэглэгчдэд өгнө. Вэб хэрэглэгчид хэлний болон бусад гадаад нөөцүүдийг ашиглан орчуулга хийнэ. Дараа нь мөн тооны вэб хэрэглэгчдээр үнэлүүлнэ.

Үүнд синсетийн хувьд үнэлгээ өгсөн вэб хэрэглэгчдийн дотоод санал нийцлийг Криппендорфийн альфагаар тооцох ба коэффициентын утга -0.1-ээс их байвал дийлэнх олонхын санал авсан үгсийг сонгон авна. Эдгээр үгсийг амжилттай хянан тохиолдуулсан гэж үзээд мэргэжилтнүүдээр онтологи нутагшуулах аргачлалын дагуу НМЦ хянан тохиолдуулагчид өгнө. Хэрэв синсетийг үнэлсэн вэб хэрэглэгчид санал нийлээгүй бол НМЦ орчуулагчид санал болгоно. Үүний дараах алхмууд мэргэжилтнүүдээр онтологи нутагшуулах аргачлалын дагуу явагдана.

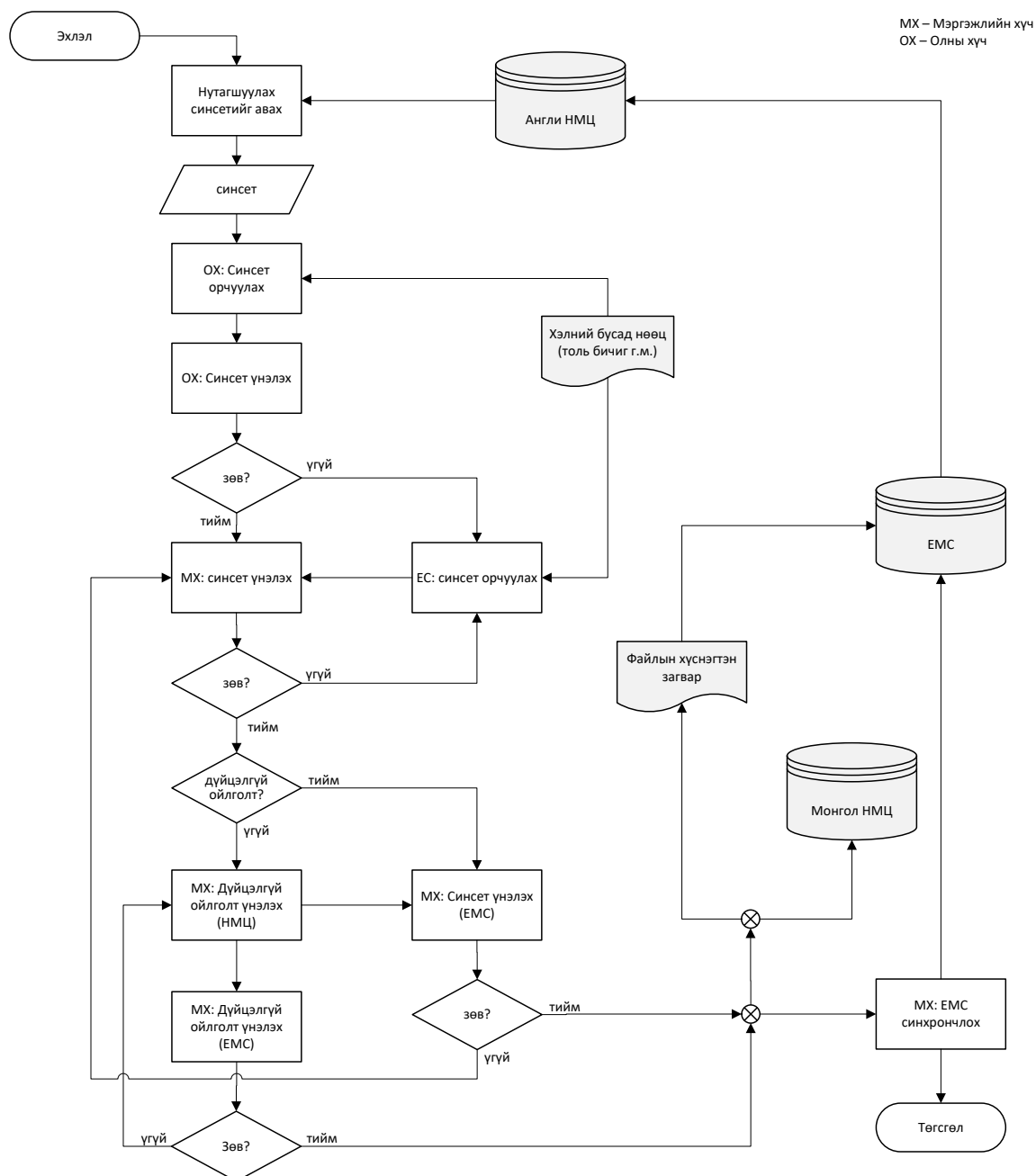
Вэб хэрэглэгчид 889 синсетийг амжилттай нутагшуулсан ба мэргэжилтнүүдийн нутагшуулсан 943 синсетээс 854-т нь давхцаж байсан.



Зураг 3.29 НМЦ-ЕМС систем бүрэлдэхүүний диаграм

Харин дүйцэлгүй ойлголтыг тодорхойлж чадаагүй. Хугацааны хувьд мэргэжилтнүүд 94 өдөр гүйцэтгэсэн бол вэб хэрэглэгчид 52 өдөр шаардсан. Энэ хоёр аргыг хослуулбал нийт барагцаагаар 57 өдөрт гүйцэтгэх боломжтой. Энэ тооцоонд (Хүснэгт 3.15) нийт амжилттай нутагшуулсан синсетийг ашигласан бөгөөд вэб хэрэглэгч болон мэргэжилтнүүд яг ижил (үгсийн хувьд яг ижил) нутагшуулсан синсетүүд гэсэн санаа биш болно.

Үүнд $943 - 889 = 54$ синсетийг вэб хэрэглэгчид нутагшуулж чадаагүй. Тэгвэл үлдсэн 54 синсетийг мэргэжилтнүүдээр нутагшуулбал түүнийг 0.43 цаг/синсетээр үржүүлэхэд нийт 42.66 цаг хуваах нь 8 цаг/өдөр буюу ойролцоогоор 5 өдөр болно. Үүн дээр вэб хэрэглэгчдийн 52 өдрийг нэмбэл нийт 57 өдөр болно. Харин зардлын хувьд 9 мэргэжилтэн 3 сар нийт 18,000.0 мянган төгрөг (нэг мэргэжилтний цалин 650.0 мян.₮) бөгөөд нэг синсетийг нутагшуулах зардал 19.0 мян.₮/синсет болно. Вэб хэрэглэгчдийн хувьд 947 синсетийн орчуулгын даалгаврыг 5 өөр оролцогчдоор, мөн 947 синсетийн үнэлгээг мөн 5 өөр оролцогчдоор гүйцэтгүүлэхэд нийт 9,470 ХОД болно.



Зураг 3.30 Олны хүчээр онтологи нутагшуулах аргачлалын алгоритм

Нэг ХОД-ыг 160 ₮ гэж үнэлбэл нийт 1,515.2 мян.₮ болно. Тэгвэл 18,000.0 -13,065.9 (1151 үгтэй 854 синсетийг нутагшуулах мэргэжилтнүүдийн зардал) = 4,934.1 болно. Үүн дээр вэб хэрэглэгчдийн зардал 1,515.2-ийг нэмбэл 6,449.3 + 1,102 (19.0 мян.₮ * үлдсэн 58 синсет) бөгөөд 7,551.3 мян.₮ болж хосолмол аргаар онтологийг нутагшуулахад нийт зардлыг барагцаалбал 58 хувиар буруулах боломжтой. Энэ бүх тооцоонд дүйцэлгүй ойлголтыг зорилгын болон эх хэл дээр тодорхойлоход шаардсан зардлыг тооцоогүй болно.

Хүснэгт 3.15 Үр дүнгийн үзүүлэлтүүдийн харьцуулалт

№	Үзүүлэлт/Арга	Мэргэжилтэн	Вэб хэрэглэгч	Хосолмол
1.	Орчуулагч	4	77	81
2.	Хянан тохиолдуулагч	5	75	80
3.	Синсет	943 (нутагшуулсан)	889 (нутагшуулж чадаагүй 58)	943 (вэб хэрэглэгчдийнхтэй давхацсан 854)
4.	Үгийн тоо	1,436 (1.52 үг/синсет)	1,813 (2.03 үг/синсет)	2627 (вэб хэрэглэгчдийнхтэй давхацсан 1151, 1.34 үг/синсет)
5.	Дүйцэлгүй ойлголт (Монгол/Англи)	42/99	0/0	42/99
6.	Хугацаа	94 өдөр (0.79 цаг/синсет)	52 өдөр (0.43 цаг/синсет)	57 өдөр
7.	Зардал	18,000.0 мян.₮	1,515.2 мян.₮	7,551.3 мян.₮ (зардлыг 58.0% бууруулав)

2013 оноос эхлэн өөрийн хөгжүүлсэн аргачлалыг UPM (Universidad Politécnica de Madrid), NUIG (National University of Ireland Galway), BU (Bielefeld University) зэрэг сургуулийн хөгжүүлсэн онтологи нутагшуулах аргачлалуудтай харьцуулав (Хүснэгт 3.16). Эдгээр аргуудыг хэрэгжүүлсэн програм хангамжийг авч ажиллуулах, ижил өгөгдөл дээр турших боломжгүй байсан тул ерөнхий үзүүлэлтээр харьцуулав. Бидний хөгжүүлсэн аргачлалын үндсэн арга нь олны хүч бөгөөд онтологийн нутагшуулах зорилтот хэлний ямар нэг нөөц шаарддаггүй учир хязгаарлагдмал нөөцтэй хэл соёлын хувьд ашиглах боломжтой. Түүнээс гадна нэмэлт хэл боловсруулалтын хэрэгсэл болон машин орчуулгын үйлчилгээ шаарддаггүй.

UPM аргачлал EuroWordNet, Wiktionary зэрэг үгийн сан, Word Sense Disambiguation, Word Sense Discovery зэрэг хэл боловсруулалтын багаж, Google Translate, Babelfish, FreeTranslation зэрэг машин орчуулгын үйлчилгээ их шаарддаг. NUIG аргачлал Europarl материалын сан, үгийн утгыг ойролцоо байдлыг (semantic similarity) үнэлэх багаж, машин орчуулгын Moses Toolkit системийг ашигладаг. BU аргачлал олон хэлний онтологийн үгийн санг үүсгэдэг M-ATOLL фрэймвөркийг ашигладаг. Мөн нутагшуулах элементийн тоогоор бусад аргачлалуудаас илүү бөгөөд дүйцэлгүй

ойлголтыг хэлний болон ойлголтын түвшинд тодорхойлох, эх болон зорилтот хэл дээр шинэ ойлголт тодорхойлох боломжийг агуулснаараа шинэлэг байна.

Хүснэгт 3.16 Онтологи нутагшуулах аргачлалын ерөнхий харьцуулалт

№	Нутагшуулах аргууд	UPM (2009) [7], [10]	NUIG (2013) [17]	BU (2016) [22]	Миний аргачлал (2013) [29], [32]
Үзүүлэлтүүд					
1.	Үндсэн техник	Хагас автомат	Автомат	Олны хүч	Олны хүч
2.	Зорилтот хэлний нөөц шаардах	Тийм	Тийм	Үгүй	Үгүй
3.	Хязгаарлагдмал нөөцтэй хэлэнд ашиглах	Боломжгүй	Боломжгүй	Боломжтой	Боломжтой
4.	Мэргэжлийн хүч шаардах	Тийм	Үгүй	Үгүй	Тийм
5.	Хэл боловсруулалтын хэрэгсэл	Ашиглана	Ашиглана	Ашиглана	Ашиглахгүй
6.	Машин орчуулгын үйлчилгээ	Ашиглана	Ашиглана	Ашиглахгүй	Ашиглахгүй
7.	Нутагшуулах элементийн тоо	2	1	1	15
8.	Ялгамж тодорхойлох	Хэлний түвшинд	Үгүй	Үгүй	Хэлний, ойлголтын түвшинд
9.	Шинэ ойлголт үүсгэх	Боломжгүй	Боломжгүй	Боломжгүй	Боломжтой
10.	Үнэлгээ хийх арга	Жишиг сан	Жишиг сан	Жишиг сан	Жишиг сан
11.	Тохирол	0.72	-	> 0.70	> 0.74

Энэ харьцуулалтаас харахад бүх аргачлал тохиролын утгыг тооцохдоо жишиг сантай харьцуулсан байна. UPM аргачлалын хувьд тохиролын утга 0.72 байсан бол вэб хэрэглэгчээр гүйцэтгүүлсэн ажлын тохирол BU аргачлалын хувьд 0.70, NUM-UNITN-ний хувьд 0.74 байсан.

Олны хүчийг ашигласан BU аргачлалыг өөрийн аргачлалтай харьцуулсныг Хүснэгт 3.17-гоос үзнэ үү. Энэ хүснэгтэд үзүүлсэн хоёр аргачлал хоёул онтологийг илэрхийлэх эх хэлнээс үгсийг зорилтот хэлэнд орчуулах, орчуулсан үгсийг үнэлэх 2 үет олны хүчний даалгаврыг тодорхойлж онтологийг нутагшуулсан. BU аргачлалын хувьд ойлголтыг илэрхийлэх үгийг өгүүлбэрт оруулж орчуулсан бол миний аргачлалаар эх хэлний синсетийг утгын тайлбарын хамт харуулж тухайн ойлголтод тохирох монгол үгийг орчуулуулж авсан.

Хүснэгт 3.17 Олны хүчээр онтологи нутагшуулах аргачлалын туршилт, үр дүнгийн харьцуулалт

№	Аргачлал		Миний аргачлал
	Үзүүлэлт	BU аргачлал	
1.	Олны хүчний платформ	CrowdFlower	CrowdCrafting
2.	Оролцогч	Вэб хэрэглэгч	Вэб хэрэглэгч
3.	Нийт оролцогчид	75	152
4.	Оролцогчдын хэл	Англи хэлтэй япон хүмүүс	Англи хэлтэй монгол хүмүүс
5.	Оролцогчдыг өдөөх арга	Хөлс төлөх	Сайн дураар оролцуулах
6.	Даалгаврын үе шат	2 (орчуулах/үнэлэх)	2 (орчуулах/үнэлэх)
7.	Нийт ХОД	10,953	9,490
8.	Нэг даалгаврын оролцогч	3	5
9.	Оролцогчдод зарцуулсан зардлын дүн	985.77 \$	0 \$
10.	Эх хэл	Англи	Англи
11.	Эх хэлний үгс	1,217	1,501
12.	Зорилтот хэл	Япон	Монгол
13.	Нутагшуулсан үгс	-	1,813
14.	Ажлын нэгтгэх арга	дийлэнх олонхийн санал	дийлэнх олонхийн санал
15.	Олны дотоод санал нийцлийг тооцох арга	байхгүй	Флайс каппа, Криппендорфын альфа
16.	Тохирол	> 0.70	> 0.74

Энэ харьцуулалтаас харахад нутагшуулсан онтологийг илэрхийлэх үгс хэр оновчтой тохирч байгааг харуулах тохирлын утга юм. Миний боловсруулсан аргачлалын хувьд 4 хувиар өссөн үр дүн гарсан болно.

ДҮГНЭЛТ

Энэ ажлаар орчуулга болон онтологийг нутагшуулахад олны хүчийг ашиглах боломжийг судалж, олон хэлний нэгдмэл онтологид тавигдах ялгамжийн төрлийг тодорхойлсны үндсэн дээр ялгамжит онтологийг нутагшуулах вэб хэрэглэгч болон мэргэжлийн олны хүчийг хослуулан ашиглах аргачлалыг боловсруулж, түүнийг хэрэгжүүлсэн НМЦ-ийг нутагшуулах системийг хөгжүүлж Монгол, Бенгал, Хятад хэлэнд амжилттай туршлаа. Энэ ажлаар дараах үр дүнд хүрлээ.

1. Онтологи нутагшуулахад олны хүчийг ашиглан өндөр чанартай үр дүнг гарган авах боломжтой гэдгийг харуулж чадлаа. Үүнд вэб хэрэглэгчид болон мэргэжлийн олны хүчийг ашигласан хоёр арга боловсруулсан.
2. Мэргэжлийн олны хүчээр онтологийг нутагшуулах явцад ойлголтын түвшинд 2 (*сацрал үгээр илэрхийлэх ойлголт, дүйцэлгүй ойлголт*), үгийн утгалбарын түвшинд 2 (*ижил нэр, дүйцэлгүй үг*) болон синсетийн түвшинд 2 (*синсетийн үгсийн зөрүү, синсетийн тайлбарын өөрчлөлт*) нийт ялгамжийн 6 төрлийг тодорхойлсон.
3. НМЦ-ийг нутагшуулах ажлын даалгавар тодорхойлж түүнд тохирсон мэргэжилтнүүдээр ойлголтыг орчуулах болон шинэ ойлголт нэмэх гэсэн 2 үндсэн алгоритмтай, мэдлэгийн эх үүсвэрийг хадгалах гарал үүслийн загвар бүхий аргачлалыг боловсруулсан. Энэ аргачлалын дагуу 985 ойлголт, 1501 үгтэй онтологийг англи хэлнээс нутагшуулж нийт 1,042 ойлголт, 1,436 үг бүхий орон зайн монгол онтологийг үүсгэж чадсан. Энэ ажлын хүрээнд англиас монгол хэлэнд 42, монголоос англи хэлэнд 99 дүйцэлгүй ойлголт тодорхойлсон бөгөөд ЕМС-г шинэ 99 ойлголтоор өргөтгөж чадлаа. Мэргэжлийн олны хүчний зардлыг бууруулахын тулд вэб хэрэглэгчдээр онтологийг нутагшуулах аргачлалыг боловсруулсан. Энэ хоёр аргачлалыг нэгтгэж вэб хэрэглэгчдээр синсетийг орчуулуулж үнэлүүлэх, гарсан үр дүнг мэргэжилтнүүдээр хянуулах хосолмол аргачлалыг гаргасан болно.
4. Онтологи нутагшуулах ажлын хүрээнд олны хүчээр гүйцэтгэсэн хувь нэмрийг нэгтгэхэд Флайс каппа, Криппендорфийн альфа зэрэг статистик хэмжүүрийг ашигласан шинэлэг туршилт боллоо. Туршилтыг хийхдээ мэргэжилтнүүдээр үнэлүүлсэн онтологийг жишиг сан болгож түүнтэй харьцуулж тохирлын утгуудыг тооцсон болно. Санал нийцлийн коэффициентын утга -0.1 дээш байхад тохирол өсөх бөгөөд энэ утгаас дээш санал нийцэлтэй синсетийн үгсээс

дийлэнх олонхын Зөв санал авсан үгсийг сонгоход тохиролын утга Альфа, Альфа-о, Каппа аргуудын хувьд харгалзан 0.738, 0.739, 0.743 байх ба 74 хувийн үнэн үр дүн гарган авч чадлаа. Φ_1 -оноо 0.81.

Цаашид энэ аргачлалд оролцогчдын ажлын гүйцэтгэлийн чанарыг тогтоож түвшинд нь тохирсон даалгаврыг оноох, шаардлага хангахгүй гүйцэтгэлтэй оролцогчдын оруулах хувь нэмрийг хязгаарлах, орчуулгын нарийвчилсан мөрдлөг боловсруулж оролцогчдыг хангах, оролцогчдын дотоод эрмэлзлийг өдөөх загвар боловсруулж хэрэгжүүлэх замаар сайжруулах боломжтой байна.

НОМ ЗҮЙ

- [1] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Sci. Am.*, vol. 284, no. 5, pp. 34–43, 2001.
- [2] N. Shadbolt, W. Hall, and T. Berners-Lee, “The semantic web revisited,” *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, 2006.
- [3] G. Antoniou and F. Van Harmelen, *A Semantic Web Primer*, 2nd ed. Cambridge, Massachusetts London, England, 2008.
- [4] C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “DBpedia : A Nucleus for a Web of Open Data,” *Nucleus*, vol. 4825, pp. 722–735, 2007.
- [5] D. Vrandečić and M. Krötzsch, “Wikidata: A Free Collaborative Knowledgebase,” *Commun. ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014.
- [6] F. M. Suchanek, G. Kasneci, and G. Weikum, “YAGO: A Large Ontology from Wikipedia and WordNet,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 6, no. 3, pp. 203–217, Sep. 2008.
- [7] M. Espinoza, E. Montiel-Ponsoda, and A. Gómez-Pérez, “Ontology localization,” in *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*, 2009, pp. 33–40.
- [8] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, “Introduction to WordNet: An On-line Lexical Database *,” *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, 1990.
- [9] K. Lindén and L. Carlson, “FinnWordNet – Finnish WordNet by Translation,” *LexicoNordica – Nord. J. Lexicogr.*, vol. 17, pp. 119–140, 2010.
- [10] M. Espinoza, A. Gómez-Pérez, and E. Mena, “LabelTranslator - A Tool to Automatically Localize an Ontology,” in *ESWC'08 Proceedings of the 5th European semantic web conference on The semantic web: research and applications*, 2008, pp. 792–796.
- [11] C. Roussey, F. Pinet, M.-A. Kang, and O. Corcho, “an Introduction to Ontologies and Ontology Engineering,” *Adv. Inf. Knowl. Process.*, vol. 1, pp. 9–38, Jul. 2011.
- [12] F. Giunchiglia, B. Dutta, and V. Maltese, “From Knowledge Organization to Knowledge Representation,” in *ISKO UK Conference*, 2013, no. June.
- [13] F. Giunchiglia, B. Dutta, and V. Maltese, “Faceted Lightweight Ontologies,” in *Conceptual Modeling Foundations and Applications*, vol. 5600, 2009, pp. 36–51.
- [14] F. Giunchiglia, M. Marchese, and I. Zaihrayeu, “Encoding Classifications into Lightweight Ontologies,” in *Journal on Data Semantics VIII*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 57–81.
- [15] M. Espinoza, A. Gómez-Pérez, and E. Mena, “Enriching an Ontology with Multilingual Information,” in *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings*, S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 333–347.
- [16] P. Cimiano, E. Montiel-ponsoda, P. Buitelaar, and M. Espinoza, “A Note on Ontology Localization,” *Appl. Ontol.*, vol. 5, pp. 127–137, 2010.

- [17] M. Arcan and P. Buitelaar, “Ontology Label Translation,” in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, 2013, pp. 40–46.
- [18] J. P. McCrae, M. Arcan, K. Asooja, J. Gracia, P. Buitelaar, and P. Cimiano, “Domain adaptation for ontology localization,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 36, pp. 23–31, 2016.
- [19] S. Walter, C. Unger, and P. Cimiano, “M-ATOLL: a framework for the lexicalization of ontologies in multiple languages,” in *International Semantic Web Conference*, 2014, pp. 472–486.
- [20] P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek, “LexInfo: A declarative model for the lexicon-ontology interface,” *J. Web Semant.*, vol. 9, no. 1, pp. 29–51, 2011.
- [21] M. Benjamin and P. Radetzky, “Multilingual Lexicography with a Focus on Less-Resourced Languages : Data Mining , Expert Input , Crowdsourcing , and Gamification Acquiring Lexical Data for LRLs,” *9th Ed. Lang. Resour. Eval. Conf.*, 2014.
- [22] B. Lanser, C. Unger, and P. Cimiano, “Crowdsourcing Ontology Lexicons,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 3477–3484.
- [23] Hasi and Nasun-Urt, “The automatic construction method of mongolian wordnet noun sets of synonyms,” in *Proceedings - 2011 4th International Conference on Intelligent Networks and Intelligent Systems, ICINIS 2011*, 2011, pp. 195–198.
- [24] A. Chagnaa, “Lexical semantic network for Mongolian,” in *Khurel Togoot*, 2010, pp. 207–210.
- [25] A. Ganbold, F. Farazi, M. Reyad, O. Nyamdavaa, and F. Giunchiglia, “Managing Language Diversity Across Cultures : the English-Mongolian Case Study,” *Int. J. Adv. Life Sci.*, vol. 6, no. 3, pp. 167–176, 2014.
- [26] A. Ganbold, F. Farazi, and G. Fausto, “Managing language diversity across cultures: the English-Mongolian case study,” in *International Conference on Knowledge Modelling and Knowledge Management (ICKM)*, 2013, pp. 1–10.
- [27] A. Ganbold, F. Farazi, and F. Giunchiglia, “An Experiment in Managing Language Diversity Across Cultures,” in *eKNOW 2014 : The Sixth International Conference on Information, Process, and Knowledge Management*, 2014, no. c, pp. 51–57.
- [28] Г. Амарсанаа and Ч. Алтангэрэл, “Онтологи нутагшуулахад олны хүчийг ашиглах туршилт,” in *Монголын мэдээллийн технологи*, 2015, pp. 1–4.
- [29] A. Ganbold and A. Chagnaa, “Crowdsourcing Localization of Ontology and Geographical Names,” in *The Eighth International Conference on Frontiers of Information Technology*, 2015, pp. 120–124.
- [30] A. Ganbold, F. Farazi, and F. Giunchiglia, “UKC Translation: Guidelines and methodology,” Trento, Italy, 2013.
- [31] A. Ganbold, F. Farazi, A. Chagnaa, and F. Giunchiglia, “UKC Translation into Mongolian,” Trento, Italy, 2014.
- [32] A. Ganbold, F. Bux, F. Farazi, I. Zaihrayeu, A. Autayeu, M. Marasca, and F. Giunchiglia, “A diversity aware methodology for a continuously evolving translation of the UKC,” Trento, Italy, 2014.

- [33] E. Bignotti, A. Ganbold, and V. Maltese, “Mind Product Etype Graph Description,” Trento, Italy, 2014.
- [34] A. P. Sheth and C. Ramakrishnan, “Semantic (Web) technology in action: Ontology driven information systems for search, integration, and analysis,” *IEEE Data Eng. Bull.*, vol. 26, no. 4, p. 40, 2003.
- [35] E. Blomqvist, “The use of Semantic Web technologies for decision support – a survey,” *Semant. Web*, vol. 5, no. 3, pp. 177–201, 2014.
- [36] T. Pellissier Tanon, D. Vrandečić, S. Schaffert, T. Steiner, and L. Pintscher, “From Freebase to Wikidata: The Great Migration,” in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 1419–1428.
- [37] K. Bollacker, R. Cook, and P. Tufts, “Freebase: A shared database of structured general human knowledge,” *Proc. Natl. Conf. Artif. Intell.*, vol. 22, no. 2, p. 1962, 2007.
- [38] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: a collaboratively created graph database for structuring human knowledge,” *SIGMOD 08 Proc. 2008 ACM SIGMOD Int. Conf. Manag. data*, pp. 1247–1250, 2008.
- [39] D. Lenat and R. Guha, “Building large knowledge-based systems: Representation and inference in the CYC project,” *Artif. Intell.*, vol. 61, no. 1, pp. 41–52, 1993.
- [40] M. Färber, B. Ell, C. Menne, A. Rettinger, and F. Bartscherer, “Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO,” *Semant. web J. Interoperability, Usability, Appl.*, vol. 7, no. 5, 2016.
- [41] F. Giunchiglia, V. Maltese, and D. Biswanath, “Domains and context: first steps towards managing diversity in knowledge,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 12–13, pp. 53–63, 2012.
- [42] F. Giunchiglia, B. Dutta, V. Maltese, and F. Farazi, “A facet-based methodology for the construction of a large-scale geospatial ontology,” *J. Data Semant.*, vol. 1, no. 1, pp. 57–73, 2012.
- [43] Н. Нансалмаа, *Үгийн сан судлал*. Улаанбаатар: Адмон, 2015.
- [44] R. Bakshi and A. Vijhni, “Semantic Web-An Extensive Literature Review,” *Int. J. Mod. Trends Eng. Res.*, vol. 2, no. 08, pp. 278–285, 2015.
- [45] F. Giunchiglia and M. Fumagalli, “From ER Models to the Entity Model,” Trento, Italy, 2014.
- [46] P. Vossen, “EuroWordNet: a multilingual database for information retrieval,” *Proc. DELOS Work. Cross-language Inf. Retr.*, pp. 5–7, 1997.
- [47] L. Bentivogli and E. Pianta, “Looking for lexical gaps,” in *In Proceedings of the Ninth EURALEX International Congress*, 2000, pp. 663–669.
- [48] J. Howe, “The Rise of Crowdsourcing,” *Wired Mag.*, vol. 14, no. 06, pp. 1–5, 2006.
- [49] A. J. Quinn and B. B. Bederson, “Human Computation: A Survey and Taxonomy of a Growing Field,” *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, pp. 1403–1412, 2011.
- [50] M. O’Hagan, *Studies on translation and multilingualism Crowdsourcing translation*, 1st ed. Publications Office of the European Union, 2012.

- [51] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide Web," *Commun. ACM*, vol. 54, no. 4, p. 86, 2011.
- [52] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell, "Collaborative workflow for crowdsourcing translation," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 2012, pp. 1191–1194.
- [53] Jonathan Ledlie, Billy Otero, Einat Minkov, and Joseph Polifroni, "Crowd translator: on building localized speech recognizers through micropayments," *ACM SIGOPS Operating Systems*, New York, NY, USA, pp. 84–89, 2010.
- [54] Omar F. Zaidan and Chris Callison-Burch, "Crowdsourcing translation: professional quality from non-professionals," in *HLT'11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011, pp. 1220–1229.
- [55] T. Aikawa, K. Yamamoto, and H. Isahara, "The impact of crowdsourcing post-editing with the collaborative translation framework," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7614 LNAI, pp. 1–10.
- [56] C. Callison-Burch, "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk," *Lang. Speech*, vol. 1, no. August, p. 286, 2009.
- [57] J. Corney, A. Lynn, C. Torres, P. Di Maio, W. Regli, G. Forbes, and L. Tobin, "Towards crowdsourcing translation tasks in library cataloguing, a pilot study," in *4th IEEE International Conference on Digital Ecosystems and Technologies - Conference Proceedings of IEEE-DEST 2010, DEST 2010*, 2010, pp. 572–577.
- [58] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," ... *40Th Annu. Meet. ...*, no. July, pp. 311–318, 2002.
- [59] Z. Li, C. Callison-Burch, C. Dyer, J. Ganitkevitch, S. Khudanpur, L. Schwartz, W. N. G. Thornton, J. Weese, and O. F. Zaidan, "Joshua: An Open Source Toolkit for Parsing-based Machine Translation," *Fourth Work. Stat. Mach. Transl.*, no. March, pp. 135–139, 2009.
- [60] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of Association for Machine Translation in the Americas*, 2006, pp. 223–231.
- [61] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder, "Findings of the 2009 Workshop on Statistical Machine Translation," in *WMT-2009*, 2009, no. March, pp. 1–28.
- [62] M. Schulze, "A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method," *Soc. Choice Welfare*, vol. 36, no. 2, pp. 267–303, 2011.
- [63] C. Spearman, "The proof and measurement of association between two things. By C. Spearman, 1904.," *Am. J. Psychol.*, vol. 100, no. 3–4, pp. 441–471, 1987.
- [64] M. Rico and C. Unger, "Lemonade: A Web Assistant for Creating and Debugging Ontology Lexica," in *Natural Language Processing and Information Systems: 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015, Proceedings, C.*

- Biemann, S. Handschuh, A. Freitas, F. Meziane, and E. Métais, Eds. Cham: Springer International Publishing, 2015, pp. 448–452.
- [65] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
 - [66] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.
 - [67] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, vol. 79. 2004.
 - [68] D. Retelny, S. Robaszkiewicz, A. To, W. S. Lasecki, J. Patel, N. Rahmati, T. Doshi, M. Valentine, and M. S. Bernstein, “Expert Crowdsourcing with Flash Teams,” *Proc. 27th Annu. ACM Symp. User interface Softw. Technol. - UIST '14*, pp. 75–85, 2014.
 - [69] M. M. R. Abdelhamid Abdelnaby, “Provenance in Open Data Entity-Centric Aggregation,” University of Trento, 2015.
 - [70] J. D. J. Deng, W. D. W. Dong, R. Socher, L.-J. L. L.-J. Li, K. L. K. Li, and L. F.-F. L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2–9, 2009.
 - [71] M. Larson, M. Soleymani, M. Eskevich, P. S. Yandex, R. R. Ordelman, and G. Jones, “The Community and the Crowd: Multimedia Benchmark Data- set Development.”

Талархал

Энэхүү докторын судалгааны ажлыг гүйцэтгэхэд үнэтэй зөвлөгөө, санал шүүмж өгсөн, тусалж дэмжсэн удирдагч дэд проф. Чагнаагийн Алтангэрэл, хамтран удирдагч проф. Паусто Жүнхилъя, эрдэм шинжилгээний ажлын зөвлөх дэд проф. Чоймаагийн Лодойравсал нартаа гүнээ талархаж байна. Мөн Италийн Трэнтогийн их сургуулийн KnowDive судалгааны багийн хамт олон, МУИС-ийн Хэрэглээний шинжлэх ухаан, инженерчлэлийн сургуулийн Машин оюуны лабораторийн гишүүддээ баярлалаа. Туршилтад оролцсон бүх хүмүүс, туршилт хийхэд тусалсан МУИС-ийн Шинжлэх ухааны сургуулийн Хүмүүнлэгийн ухааны салбарын багш нарт талархаж байна.

Эцэст нь, үргэлж урам зориг, ухаанаар тэтгэж бүх талаар тусалж дэмжсэн миний хайртай аав ээж, эхнэр, хүүхдүүд, ах, дүү нартаа маш их баярлалаа.

ХАВСРАЛТ А. ЕРТӨНЦИЙН МЭДЛЭГИЙН САНГИЙН ХОЛБООС

№	Холбоосууд	
1.	is-a	Утгазүйн холбоос (Ойлголт хоорондын)
2.	has-aspect	
3.	is-agent	
4.	part-of	
5.	value-of	
6.	substance-of	
7.	member-of	
8.	metaphor-of	
9.	antonym	Үгзүйн холбоос (Утгалбар хоорондын)
10.	pertains-to	
11.	pertainym-of	
12.	participle-of-verb	
13.	also-see	
14.	related-form	
15.	homograph-of	
16.	lexical entailment	Үг-утгазүйн холбоос (Синсет хоорондын)
17.	troponymy	
18.	cause-of	
19.	similar-to	
20.	verb-group	

ХАВСРАЛТ Б. ФАЙЛЫН ХҮСНЭГТЭН ЗАГВАРЫН ЖИШЭЭ

Senses (Утгалбар)

Cased Word Lemma	Word Forms	Concept UK ID	Word Sense Rank	Concept Word Rank	PoS	Description	Operation	Language	Reference	Note	Reference_type
hovel		19379	1	1	n	small crude shelter used as a dwelling	UPDATE	en			USER
hut		19379	2	2	n	small crude shelter used as a dwelling	UPDATE	en			USER
hutch		19379	2	3	n	small crude shelter used as a dwelling	UPDATE	en			USER
shack		19379	1	4	n	small crude shelter used as a dwelling	UPDATE	en			USER
shanty		19379	1	5	n	small crude shelter used as a dwelling	UPDATE	en			USER
дэглэй		19379	1	4	n	хүн амьдарч болохуйц жижиг орогнох байр; урц	ADD	mn	Baasantsend Ochirjav (baask0408@yahoo.com)	OLM	USER
жолум		19379	1	5	n	хүн амьдарч болохуйц жижиг орогнох байр; урц	ADD	mn	Baasantsend Ochirjav (baask0408@yahoo.com)	OLM	USER
овоохой		19379	1	1	n	хүн амьдарч болохуйц жижиг орогнох байр; урц	ADD	mn	Baasantsend Ochirjav (baask0408@yahoo.com)	OLM	USER
оромж		19379	1	3	n	хүн амьдарч болохуйц жижиг орогнох байр; урц	ADD	mn	Baasantsend Ochirjav (baask0408@yahoo.com)	OLM	USER
өвөн		19379	1	6	n	хүн амьдарч болохуйц жижиг орогнох байр; урц	ADD	mn	Baasantsend Ochirjav (baask0408@yahoo.com)	OLM	USER
өөвөн		19379	1	7	n	хүн амьдарч болохуйц жижиг орогнох байр; урц	ADD	mn	Baasantsend Ochirjav (baask0408@yahoo.com)	OLM	USER
тов		19379	1	8	n	хүн амьдарч болохуйц жижиг орогнох байр; урц	ADD	mn	Baasantsend Ochirjav (baask0408@yahoo.com)	OLM	USER
урц		19379	1	2	n	хүн амьдарч болохуйц жижиг орогнох байр; урц	ADD	mn	Baasantsend Ochirjav (baask0408@yahoo.com)	OLM	USER

* OLM – Ontology Localization in Mongolian

Relations (Холбоос)

Parent Concept UK ID	Parent Concept Label	Child Concept UK ID	Child Concept Label	Relation Kind	Operation	Language	Reference	Note	Reference_type
19379	Hut	-40	acap	IS_A	ADD	mn	Oyundari Nyamdavaa (oyundari.n@gmail.com)	Ontology Localization in Mongolian	USER

19379	Hut	-46	жовгон	IS_A	ADD	mn	Oyundari Nyamdavaa (oyundari.n@gmail.com)	Ontology Localization in Mongolian	USER
19379	Hut	-71	отог	IS_A	ADD	mn	Oyundari Nyamdavaa (oyundari.n@gmail.com)	Ontology Localization in Mongolian	USER
19369	House	19379	Hut	IS_A	UPDATE	en			USER

Gaps (Дүйцэлгүй ойлголт)

Cased Word Lemma	Concept UK ID	Language	Operation	Literal Translation	Reference	Note	Reference_type
майхан	-40	en	ADD	a shelter made of cloth which is supported by poles; in many different shapes	Baasantsend Ochirjav (baask0408@yahoo.com)	Ontology Localization in Mongolian	USER
асар	-40	en	ADD	a shelter made of cloth which is supported by poles; in many different shapes	Baasantsend Ochirjav (baask0408@yahoo.com)	Ontology Localization in Mongolian	USER
жодгор	-40	en	ADD	a shelter made of cloth which is supported by poles; in many different shapes	Baasantsend Ochirjav (baask0408@yahoo.com)	Ontology Localization in Mongolian	USER
жовгон	-46	en	ADD	kind of shelter	Baasantsend Ochirjav (baask0408@yahoo.com)	Ontology Localization in Mongolian	USER
жогвон	-46	en	ADD	kind of shelter	Baasantsend Ochirjav (baask0408@yahoo.com)	Ontology Localization in Mongolian	USER
отог	-71	en	ADD	a lodge or shelter for hunters and tourists etc.	Khulan Tsog-Erdene (khulan_0726@yahoo.com)	Ontology Localization in Mongolian	USER

ХАВСРАЛТ В. ХЭЛНИЙ ЦӨМИЙГ ЭКСПОРТЛОХ RDF ФАЙЛЫН ЖИШЭЭ

RDF тодорхойлолт

Класс	Тайлбар
wn20schema:NounSynset	A synset including noun word senses.
wn20schema:VerbSynset	A synset including verb word senses.
wn20schema:AdjectiveSynset	A synset including adjective word senses.
wn20schema:AdverbSynset	A synset including adverb word senses.
Шинж	Тайлбар
uk:sense	A container for sense
wn20schema:senseLabel	A property filled with the values of the lexicalForms of all the Words in a Synset
wn20schema:lexicalForm	A datatype relation between Word and its lexical form.
wn20schema:gloss	It specifies the gloss for a synset.
uk:wordSenseRank	It defines the position of the sense of word
uk:synsetWordRank	It defines the position of the word in synset
uk:gap	It specifies a gap in a language
uk:operation	It determines the operation on elements
xml:lang	It defines the language of elements
wn20schema:hyponymOf	It specifies that the child concept is a hypernym of the parent concept.
uk:hasAspect	It specifies that the parent concept has secondary aspect of the child concept
uk:isAgent	It specifies that the parent concept has secondary aspect of the child concept. But the child concept must be agentive.
wn20schema:partMeronymOf	It specifies that the child concept is a part of the parent concept.
wn20schema:entails	It specifies that the child concept is an entailment of the parent concept.
wn20schema:attribute	It defines the value-of relation between concepts in which the child concept is a value of the parent concept.
wn20schema:memberMeronymOf	It specifies that the child concept is a member of the parent concept.
wn20schema:substanceMeronymOf	It specifies that the child concept has substance of the parent concept.
wn20schema:similarTo	It specifies that the child concept is similar in meaning to the parent concept.
wn20schema:causes	It defines causal relations between concepts.
uk:pertainymOf	It specifies adjectival relations between concepts.
wn20schema:sameVerbGroupAs	It defines verbal concepts in similar meaning
uk:isMetaphor	It defines concepts with metaphoric meaning.

WordNet 2.0-ийн RDF Schema-ийг <http://www.w3.org/2006/03/wn/wn20/download>

хаягаас татаж авч болно.

RDF баримтын жишээ

RDF баримт нь хүснэгтэн талбар бүрт үүсдэг.

Senses талбарын RDF баримтын жишээ

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE rdf:RDF [
  <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
```

```

<!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#>
<!ENTITY wn20instances 'http://www.w3.org/2006/03/wn/wn20/instances/'>
<!ENTITY wn20schema 'http://www.w3.org/2006/03/wn/wn20/schema/'>
<!ENTITY uk 'http://uk.disi.unitn.it/ukc/'>
]>
<rdf:RDF
  xmlns:rdf="&rdf;"
  xmlns:rdfs="&rdfs;"
  xmlns:wn20instances="&wn20instances;"
  xmlns:wn20schema="&wn20schema;"
  xmlns:uk="&uk;">
<rdf:Description rdf:about="&uk;concepts/16038">
  <wn20schema:NounSynset xml:lang="en" rdf:parseType="Literal">
    <wn20schema:gloss>a portable container for carrying several objects;
    &quot;the musicians left their instrument cases
    backstage&quot;</wn20schema:gloss>
    <uk:sense>
      <wn20schema:senseLabel>case</wn20schema:senseLabel>
      <uk:wordSenseRank>9</uk:wordSenseRank>
      <uk:synsetWordRank>1</uk:synsetWordRank>
    </uk:sense>
    <uk:operation>UPDATE</uk:operation>
  </wn20schema:NounSynset>
  <wn20schema:NounSynset xml:lang="it" rdf:parseType="Literal">
    <wn20schema:gloss>un contenitore usato per riporre in modo permanente, o
    per trasportare, materiali solidi di qualsiasi genere</wn20schema:gloss>
    <uk:sense>
      <wn20schema:senseLabel>scatola</wn20schema:senseLabel>
      <uk:wordSenseRank>1</uk:wordSenseRank>
      <uk:synsetWordRank>1</uk:synsetWordRank>
    </uk:sense>
    <uk:operation>ADD</uk:operation>
  </wn20schema:NounSynset>
  <wn20schema:NounSynset xml:lang="mn" rdf:parseType="Literal">
    <wn20schema:gloss>эд зүйл зөөврийн авсаархан сав</wn20schema:gloss>
    <uk:sense>
      <wn20schema:senseLabel>хайрцар сав</wn20schema:senseLabel>
      <uk:wordSenseRank>1</uk:wordSenseRank>
      <uk:synsetWordRank>1</uk:synsetWordRank>
    </uk:sense>
    <uk:operation>ADD</uk:operation>
  </wn20schema:NounSynset>
</rdf:Description>
</rdf:RDF>

```

relations талбарын RDF баримт

```

<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE rdf:RDF [
  <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#>
  <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#>
  <!ENTITY wn20instances 'http://www.w3.org/2006/03/wn/wn20/instances/'>
  <!ENTITY wn20schema 'http://www.w3.org/2006/03/wn/wn20/schema/'>
  <!ENTITY uk 'http://uk.disi.unitn.it/ukc/'>
]>
<rdf:RDF
  xmlns:rdf="&rdf;"
  xmlns:rdfs="&rdfs;"
  xmlns:wn20instances="&wn20instances;"
  xmlns:wn20schema="&wn20schema;"
  xmlns:uk="&uk;">
<!-- pillbox IS_A case -->
<rdf:Description rdf:about="&uk;concepts/16038">
  <wn20schema:hyponymOf rdf:resource="&uk;concepts/21672" />
  <uk:operation>ADD</uk:operation>
</rdf:Description>
<!-- television series IS_A serial -->

```

```

<rdf:Description rdf:about="&uk;concepts/35489">
  <wn20schema:hyponymOf rdf:resource="&uk;concept/120705" />
  <uk:operation>ADD</uk:operation>
</rdf:Description>
</rdf:RDF>

```

gaps талбарын RDF баримт

```

<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE rdf:RDF [
  <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
  <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#'>
  <!ENTITY wn20instances 'http://www.w3.org/2006/03/wn/wn20/instances/'>
  <!ENTITY wn20schema 'http://www.w3.org/2006/03/wn/wn20/schema/'>
  <!ENTITY uk 'http://uk.disi.unitn.it/ukc/'>
]>

<rdf:RDF
  xmlns:rdf="&rdf;"
  xmlns:rdfs="&rdfs;"
  xmlns:wn20instances="&wn20instances;"
  xmlns:wn20schema="&wn20schema;"
  xmlns:uk="&uk;">
  <rdf:Description rdf:about="&uk;concepts/120709">
    <uk:gap xml:lang="mn"/>
    <uk:operation>ADD</uk:operation>
  </rdf:Description>
</rdf:RDF>

```

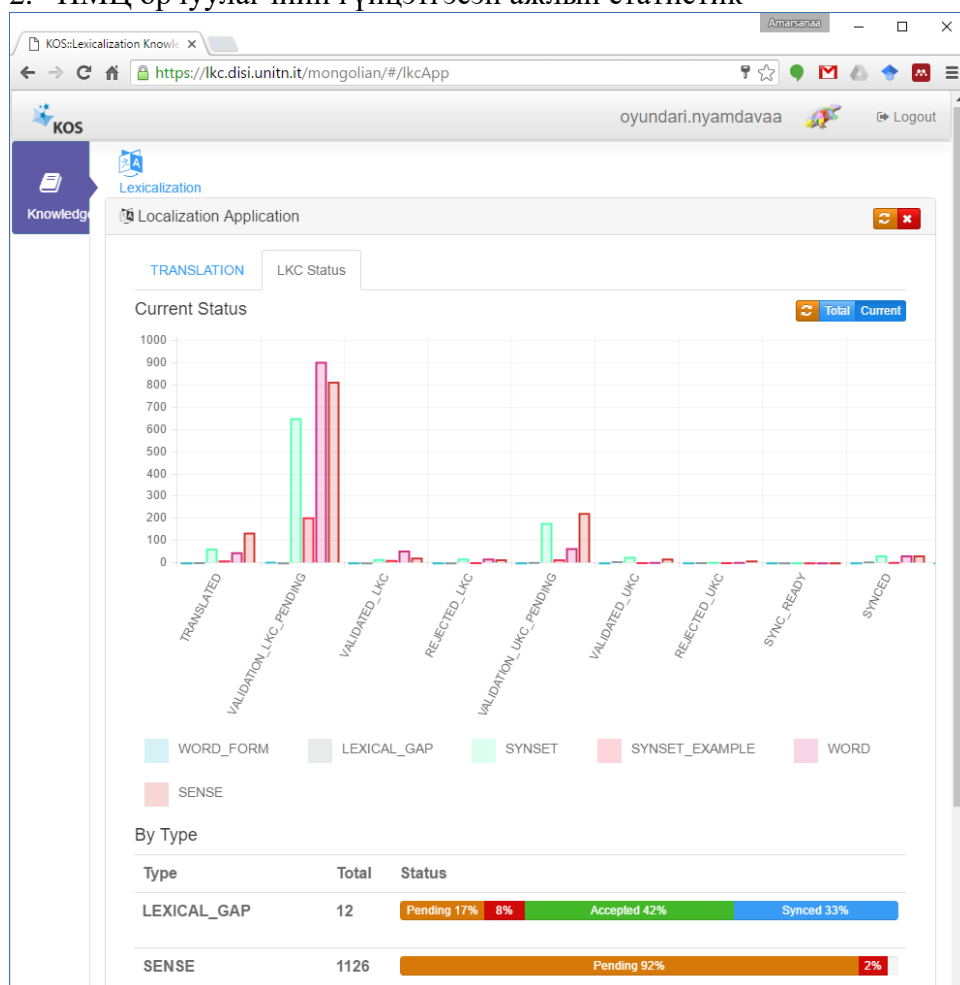
ХАВСРАЛТ Г. НМЦ-ийг нутагшуулах системийн хэрэглэгчийн зарим интерфейс

1. НМЦ орчуулагчид оноосон дэд моднууд

The screenshot shows the 'Localization Application' interface. It includes a sidebar with 'Lexicalization' and 'Knowledge' tabs. The main content area has a 'TRANSLATION' tab and a 'LKC Status' link. Below this is a table with columns: AssignmentId, Root Concept, Total Concepts, Quality, and Status. There are also 'Statistics', 'Translate', and 'Statistics' buttons for each row.

AssignmentId	Root Concept	Total Concepts	Quality	Status
201	jungle	0/2		CLOSED
4	location	0/1227		ACCEPTED

2. НМЦ орчуулагчийн гүйцэтгэсэн ажлын статистик



3. ЕМС хянан тохиолдуулагч элементүүд үнэлэх интерфейс

KOS:Lexicalization Knowl...
Amarsanaa
https://lkc.disi.unitn.it/mongolian/#/lkcApp/ukcvalidate/1/46065

Lexicalization
Knowledge
Localization Application
Back to dashboard

46065
Update

Reference Language
English Provenance

ConceptId 46065

Gloss a domed rock formation where a core of rock has moved upward and pierced through the more brittle overlying strata

POS NOUN

Senses

Rank	Lemma	Exceptional forms
1	diapir	

Examples
Missing Example

Target Language: Mongolian
LOG
Validate Next

Lexical Gap

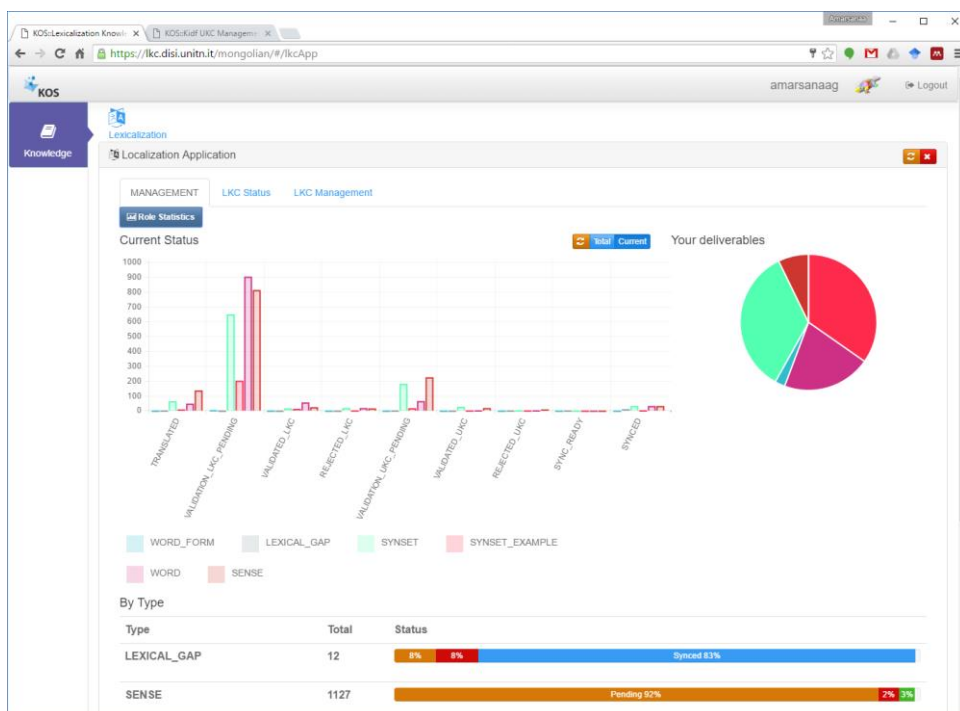
Gloss Чулууны цөм нь дээшээ өргөгдөж дээр нь байгаа илүү хэврэг давхаргийг цоолон гарч ирж буй дугуй хэлбэртэй хад чулуу

Comment

Note...

Accept Reject Validate Later

Save Accept all Reject all Submit for Synchronization



4. EMC хянан тохиолдуулагч оролцогчид дэд мод оноох интерфейс

KOS

amarsanaag Logout

Lexicalization

Localization Application

MANAGEMENT LKC Status LKC Management

New items are ready to be submitted for UKC synchronization [Submit for Synchronization](#)

User	Type	Root Concept	Total Concepts	Quality	Status	
buyannemekh.bai	TRANSLATION	body of water	0/76		ACCEPTED	Close
zolzaya.bayarsa	VALIDATION	body of water	0/76		ACCEPTED	Close
gankhuyag.badam	TRANSLATION	geological formation	0/173		ACCEPTED	Close
buyannemekh.bai	VALIDATION	geological formation	0/173		ACCEPTED	Close
zolzaya.bayarsa	TRANSLATION	relation	0/1614		ACCEPTED	Close
gankhuyag.badam	VALIDATION	relation	0/1614		ACCEPTED	Close
altangerel.chagnaa	VALIDATION	facility	7/134	100%	ACCEPTED	Close
buyannemekh.bai	VALIDATION	facility	5/134	100%	ACCEPTED	Close
gankhuyag.badam	VALIDATION	facility	14/134	93%	ACCEPTED	Close
gankhuyag.badam	TRANSLATION	facility	0/134		ACCEPTED	Close

1 2

5. EMC хянан тохиолдуулагч дүйцэлгүй ойлголтыг үнэлэх интерфэйс

KOS

amarsanaag Logout

Lexicalization

Localization Application

[Back to dashboard](#)

46274 [Update](#)

Reference Language

English [Provenance](#)

ConceptId

46274

Gloss

a block of the earth's crust bounded by faults and shifted to form peaks of a mountain range

POS

NOUN

Senses

Rank	Lemma	Exceptional forms
1	massif	

Examples

Missing Example

Target Language: Mongolian [LOG](#) [Validate Next](#)

Lexical Gap

Gloss

Уул нуруудын оргилийг үүсгэдэг шилжилт хөдөлгөөн болон эвдрэлээс болж зааглагдсан дэлхийн гадаргын царцдас

Comment

Note...

[Accept](#) [Reject](#) [Validate Later](#)

[Save](#) [Accept all](#) [Reject all](#) [Submit for Synchronization](#)

6. Гарал үүслийн бүртгэлийг мэдээллэх интерфeйс

KOS::Lexicalization Knowledge X

https://lkc.disi.unitn.it/mongolian/#/lkcApp/validate/209/24394

Amarsanaa

Log for sense

	Type	Status	Comment
Jun 24, 2016 2:33:13 by amarsanaag	SENSE	REJECTED_UKC	тээврийн сүлжээ тээврийн систем
Jun 24, 2016 2:33:13 by amarsanaag	SYNSET	REJECTED_UKC	зорчигч эсвэл бараа тээвэрлэхэд шаардлагатай тоног төхөөрөмж хэрэгслүүдээс бүрдсэн байгууламж
Jun 23, 2016 1:57:31 by buyannemekh.bai	SENSE	VALIDATION_UKC_PENDING	Submitting translation for UKC validation
Jun 23, 2016 1:57:31 by buyannemekh.bai	SYNSET	VALIDATION_UKC_PENDING	Submitting translation for UKC validation
Jun 23, 2016 1:57:28 by buyannemekh.bai	SENSE	VALIDATED_LKC	
Jun 23, 2016 1:57:28 by buyannemekh.bai	SYNSET	VALIDATED_LKC	
Nov 18, 2014 8:39:08 by amarsanaag	SENSE	VALIDATION_LKC_PENDING	[AUTO GENERATED ENTRY]: The object must be validated by an LKC validator

LOG Validate Next M

ийн цогцолбор

forms

to UKC Validation

ХАВСРАЛТ Д. ТОХИРЛЫН ҮЗҮҮЛЭЛТҮҮДИЙГ ТООЦООЛОХ R СКРИПТ

```
#The measures on crowdsourcing:
true_positive <- length(which(frameOne$ToF == T));
precision <- true_positive / nrow(frameOne);
recall <- true_positive / length(which(frameOne$ToF == T));
#CDF:
sub_set <- data.frame();precisionset_ana <- c();
recallset_ana <- c();fms_ana <- c();precisionset_a <- c();
recallset_a <- c();fms_a <- c();precisionset_k <- c();
recallset_k <- c();fms_k <- c();temp_prec <- 0;
temp_rec <- 0;truepos <- c();checker <- logical();
find_tp <- function(dataset){
  tp <- length(which(dataset$ToF == T));
  tp;}
precision_of_set <- function(dataset, tp){
  pr <- tp / nrow(dataset);
  pr;}
recall_of_set <- function(dataset, tp){
  rl <- tp / length(which(frameOne$ToF == TRUE));
  rl;}
for(i in seq(from = -1, to = 1, by = 0.1)){
  cdf_uk <- alpha_data$uk_id[which(alpha_data$alpha_na > i)];
  for(j in cdf_uk){
    sub_set <- rbind(sub_set, frameOne[which(frameOne$uk_id == j), ]);
  }
  truepos <- find_tp(sub_set);
  temp_prec <- precision_of_set(sub_set, truepos);
  temp_rec <- recall_of_set(sub_set, truepos);
  precisionset_a <- c(precisionset_a, temp_prec);
  recallset_a <- c(recallset_a, temp_rec);
  fms_a <- c(fms_a, 2*temp_rec*temp_prec / (temp_rec+temp_prec));
  sub_set <- data.frame();
}
for(i in seq(from = -1, to = 1, by = 0.1)){
  cdf_uk <- alpha_data$uk_id[which(alpha_data$alpha > i)];
  for(j in cdf_uk){
    sub_set <- rbind(sub_set, frameOne[which(frameOne$uk_id == j), ]);
  }
  truepos <- find_tp(sub_set);
  temp_prec <- precision_of_set(sub_set, truepos);
  temp_rec <- recall_of_set(sub_set, truepos);
  precisionset_ana <- c(precisionset_ana, temp_prec);
  recallset_ana <- c(recallset_ana, temp_rec);
  fms_ana <- c(fms_ana, 2*temp_rec*temp_prec / (temp_rec + temp_prec));
  sub_set <- data.frame();
}
for(i in seq(from = -1, to = 1, by = 0.1)){
  cdf_uk <- kappa_general$uk_id[which(kappa_general$kappa > i)];
  for(j in cdf_uk){
    sub_set <- rbind(sub_set, frameOne[which(frameOne$uk_id == j), ]);
  }
  truepos <- find_tp(sub_set);
  temp_prec <- precision_of_set(sub_set, truepos);
  temp_rec <- recall_of_set(sub_set, truepos);
  precisionset_k <- c(precisionset_k, temp_prec);
  recallset_k <- c(recallset_k, temp_rec);
  fms_k <- c(fms_k, 2*temp_rec*temp_prec / (temp_rec + temp_prec));
  sub_set <- data.frame();
}
#CDF data formation:
alphaseq <- seq(-1, 1, by = 0.1);
precisionset_a <- data.frame(alphaseq, precisionset_a, 'precision',
```



```

      'alpha');
recallset_a <- data.frame(alphaseq, recallset_a, 'recall', 'alpha');
fms_a <- data.frame(alphaseq, fms_a, 'f_measure', 'alpha');
precisionset_ana <- data.frame(alphaseq, precisionset_ana, 'precision',
      'alpha/-na/');
recallset_ana <- data.frame(alphaseq, recallset_ana, 'recall', 'alpha/-
      na/');
fms_ana <- data.frame(alphaseq, fms_ana, 'f_measure', 'alpha/-na/');
precisionset_k <- data.frame(alphaseq, precisionset_k, 'precision',
      'kappa');
recallset_k <- data.frame(alphaseq, recallset_k, 'recall', 'kappa');
fms_k <- data.frame(alphaseq, fms_k, 'f_measure', 'kappa');
names(precisionset_a) <- c("value", "measure", "argument", "metric");
names(recallset_a) <- c("value", "measure", "argument", "metric");
names(fms_a) <- c("value", "measure", "argument", "metric");
names(precisionset_ana) <- c("value", "measure", "argument", "metric");
names(recallset_ana) <- c("value", "measure", "argument", "metric");
names(fms_ana) <- c("value", "measure", "argument", "metric");
names(precisionset_k) <- c("value", "measure", "argument", "metric");
names(recallset_k) <- c("value", "measure", "argument", "metric");
names(fms_k) <- c("value", "measure", "argument", "metric");
recallset <- rbind(recallset_a, recallset_ana, recallset_k);
precisionset <- rbind(precisionset_a, precisionset_ana, precisionset_k);
fms <- rbind(fms_a, fms_ana, fms_k);
merged_cdf <- rbind(precisionset_a, recallset_a, fms_a, precisionset_ana
      , recallset_ana, fms_ana, precisionset_k
      , recallset_k, fms_k);
merged_cdf <- merged_cdf[-c(which(merged_cdf$value < -0.5 |
      merged_cdf$value == 1.0)), ];
#CDF data plotting /with ggplot2/:
precisionset <- precisionset[-c(which(precisionset$value < -0.5 |
      precisionset$value == 1.0)), ];
recallset <- recallset[-c(which(recallset$value < -0.5 | recallset$value
      == 1.0)), ];
fms <- fms[-c(which(fms$value < -0.5 | fms$value == 1.0)), ];
ggplot() +
  geom_line(data = recallset, aes(x = value, y = measure, colour =
      metric), linetype = "longdash", alpha = 0.7) +
  geom_line(data = precisionset, aes(x = value, y = measure, colour =
      metric), alpha = 0.7) +
  geom_line(data = fms, aes(x = value, y = measure, colour = metric),
      linetype = "dashed", alpha = 0.7) +
  geom_point(data = merged_cdf, aes(x = value, y = measure, shape =
      metric, colour = metric), size = 1.5, alpha = 0.6) +
  scale_x_continuous(name = "Metric", breaks = seq(-0.5,
      0.9, by = 0.1)) +
  scale_y_continuous(name = "Argument", breaks = seq(0, 1, by = 0.1))
  + ggtitle("CDF(X>=Metric)");

```

ХАВСРАЛТ Е. ДҮЙЦЭЛГҮЙ ОЙЛГОЛТЫН ЖИШЭЭ

Senses (Утгалбар)

Cased Word Lemma	Word Forms	Concept UK ID	Word Sense Rank	Concept Word Rank	Po S	Description
бүрх		-89	1	1	n	монгол гэрийн хэлбэртэйгээр хийж шавардсан гэр буюу отор аяны хүмүүсийн хэрэглэх оромж; "тэр бүрх барьж 3 жил суув"
атар		-86	1	1	n	хүн, малын бараг яваагүй, хагалж сэндлээгүй сэргэг шинэ нутаг; "тэд атар газар малаа отоглуулахаар явсан"
гэрүү		-81	1	2	n	уулын хүүшид ургасан шугуй
хэрүү		-81	1	1	n	уулын хүүшид ургасан шугуй
гээг		-65	1	1	n	уул толгодын ар шил; "хонь уулын гээг дээр идээшиж байгаа байх"
өвөлжөө		-62	1	1	n	өвлийн улирлыг өнгөртөл нутаглах хашаа хороо бүхий буурь, бууц; "тэднийх өвөлжөөндөө байгаа"
гэр		-60	1	1	n	нүүдэлчин Монгол үндэстний амьдрах байр; "эцэг эх нь залуу хосуудад шинэ гэр төхөөрлөө"
дөл		-56	1	1	n	уулын хэвцгий, усны хөвөөн дэхь тэгш налуу газар; "үдээс хойш тэр малаа дөл дээр авч ирсэн"
хэц		-37	1	1	n	уул хадны орой нуруу буюу цавчим бэрхлэг өндөр газар; "хонь ямааны хээл хүндэрсэн үед уул хярын хэц өөд авируулах хэрэггүй"
аранга		-32	1	1	n	бөөгийн хүүр оршуулсан газар (бөөгийн аранга); "бөөгийн аранга Монгол нутагт олон байдаг"
малын бууц		-25	1	1	n	мал хэвтэж хоноглодог газар; малын ялгадсаар дүүрсэн; "малын бууц боом өвчин тараах нян тээж байдаг"
малчны хот		-22	1	1	n	малчин айлын бууриа сэлгэж амьдардаг газар; "чоно өчигдрийн борооноор малчны хотонд оржээ"
хүүш		-6	1	1	n	уул хадны ар буюу хажуугийн нар тусахгүй сүүдэрлэсэн газар; "хүүш газар үргэлж сэрүүхэн байдаг"

Relations (Холбоос)

Parent Concept UK ID	Parent Concept Label	Child Concept UK ID	Child Concept Label	Relation Kind
49901	Mountain	-6	хүүш	PART_OF
46547	site	-22	малчны хот	IS_A
-22	малчны хот	-25	малын бууц	PART_OF
118151	Religious facility	-32	аранга	IS_A
50150	Ridge	-37	хэц	IS_A

50286	Slope	-56	дөл	IS_A
17703	Dwelling	-60	гэр	IS_A
17982	Facility	-62	өвөлжөө	IS_A
50150	Ridge	-65	гээг	IS_A
45606	forest	-81	хэрүү	IS_A
45606	forest	-81	гэрүү	IS_A
92260	arable land	-86	атар	IS_A
17703	Dwelling	-89	бүрх	IS_A

Gaps (Дүйцэлгүй ойлголт)

Cased Word Lemma	Concept UK ID	Language	Operation	Literal Translation
хүүш	-6	en	ADD	(huush) shaded place, place where the sun strikes late and disappears early; "he has been lived in a huush"
малчны хот	-22	en	ADD	(malchny khot) temporary place for herder to live; "yesterday, while it was raining, wolves entered into malchny khot"
малын бууц	-25	en	ADD	(malyn buuts) a place for livestock which is filled by droppings; "bacteria of cow-pox is in malyn buuts"
аранга	-32	en	ADD	(aranga) burial place of a shaman; "there are many aranga in Mongolia"
хэц	-37	en	ADD	(hets) mountain ridge or very steep high place; "don't push livestock with fetus near birth to go to hets"
дөл	-56	en	ADD	(dol) flat land of mountain slope or bank of river; "he brought back livestock on dol"
гэр	-60	en	ADD	(ger) a dwelling structure used by Mongolians as their home; "parents have prepared a ger for marrying couple"
өвөлжөө	-62	en	ADD	(ovoljoo) a place where the people or an animal such as sheep or goats which spend the winter in a state warmly and safely especially in Mongolia; "they are in their ovoljoo"
гээг	-65	en	ADD	(gezeg) northern ridge of a mountain; "sheeps are maybe in gezeg"
хэрүү	-81	en	ADD	(kheruu) a forest that grow on the shady side of the mountain
гэрүү	-81	en	ADD	(geruu) a forest that grow on the shady side of the mountain
атар	-86	en	ADD	(atar) the land that has not been sown or the soil that has not been worked area; "they went to atar with their livestock for fast feeding"
бүрх	-89	en	ADD	(burkh) plastered dwelling which has a shape of Mongolian traditional ger (travellers use this as a shelter); "he has built a burkh then he has been lived for 3 years"

NATIONAL UNIVERSITY OF MONGOLIA
SCHOOL OF ENGINEERING AND APPLIED SCIENCES

AMARSANAA Ganbold

**A METHODOLOGY FOR DIVERSITY-AWARE
MULTILINGUAL ONTOLOGY LOCALIZATION VIA
CROWDSOURCING**

PhD Dissertation

Scientific supervisor:

Assoc. Prof., PhD, Altangerel Chagnaa, National University of Mongolia

Co-supervisor:

Prof., Fausto Giunchiglia, University of Trento

Advisors:

Assoc. Prof., PhD Lodoiravsal Choimaa, National University of Mongolia

Prof., PhD Erdenebaatar Altangerel, Mongolian University of Science and Technology

Ulaanbaatar, 2016

ABSTRACT

To build a true, flourishing and successful Semantic Web [1]–[3] we need integrated multilingual ontologies which capture diversity across all languages and cultures in the world and, consist of thousands of concepts representing real-world entities. Developing ontologies from scratch appears to be very expensive in terms of cost and time required and often such efforts remain unfinished for decades. Ontology localization [7], [16] through translation seems to be a promising approach towards addressing the issue as it enables the greater reuse of the ontological structure and building ontologies for less-resourced languages, such as Mongolian. However, several automated approaches [7], [10], [17] for the localization activity have been developed that require a number of language resources (multilingual dictionary, thesauri, parallel corpus etc.) and NLP tools (Word sense disambiguation, machine translation etc.), managing language diversity across cultures during the ontology localization still remains as a challenge that has to be taken into account in terms of lexicalization of concepts in different languages, as well as right level of expertise.

The purpose of this research is to develop a multilingual ontology localization methodology handling diversity in lexical and conceptual level by using crowdsourcing technique and to evaluate. The following sub-objectives are defined:

1. Studying the possibility of crowdsourcing for ontology localization
2. Defining the types of diversity
3. Developing a diversity-aware ontology localization methodology
4. Evaluating the methodology

In this paper, we define a methodology for ontology localization that uses two crowdsourcing techniques based on linguistic and domain experts and web users. Crowdsourcing is a potential approach towards building ontologies with human-level accuracy as good as the well-known resources such as the Princeton WordNet [8] and FinnWordNet [9], which are built manually to obtain better quality. The main underlying idea of this methodology is to take a concept from an existing ontology in a source language and to produce the corresponding representation in a new ontology in a target language. The process includes the translation of the synset words and glosses representing the concept. A direct translation of them is provided whenever possible that could be done by web users, and some concepts are unknown or cannot be represented in a particular culture. These kind

of concepts are hard to be defined by web users but can be solved by linguistic and domain experts.

RELATED WORK

The ontology localization activity described in [7] is an attempt to address the localization issues based on their proposed guidelines and methodology for enriching with multilingual information. However, the proposed approach differs from theirs with respect to the target language and the development approach which includes concept generation in source and target language.

The ontology label translation [17] was developed to provide statistical machine translation system with additional information. However, automatically translated multilingual terms often suffer from quality issues, whereas we obtained human-level accuracy.

The preliminary crowdsourcing model for less-resourced language [21] was designated to develop lexicon with the help of Internet users. They concluded that extensive manipulation and human review by language experts in order to obtain high-quality linguistic data and capture great diversity of knowledge.

The two-stage workflow for crowdsourcing ontology lexicon through translation tasks [22] was carried out and the results were found that generating ontology lexicon via crowdsourcing was feasible. Our approach is different with respect to combining expert sourcing and crowdsourcing and number of elements being localized, while they only have ontology label translation.

The thesis consists of 3 chapters, conclusion, bibliography, 6 appendices, 23 tables and 42 figures in the total of 134 pages, and is organized as follows. Chapter 1 gives an overview of the ontology localization and diversity problems. In Chapter 2 we describe the state-of-the-art crowdsourcing translations and evaluation methods for quality control. Chapter 3 discusses our proposed methodology and reports the experimental results.

CHAPTER 1. ONTOLOGY AND DIVERSITY PROBLEM

The Universal Knowledge Core (UKC) [41] is a large-scale ontology, under development at the University of Trento, which includes hundreds of thousands of concepts (e.g., river, city) of the real-world entities (e.g., Tuul, Ulaanbaatar) and consists of three main components: *domain core*, *concept core* and *natural language core* (See Fig. 1.5). The concept core is a formal ontology in formal language DL (Description Logic) which includes concepts and

semantic relations between them. Natural language core consists of a set of languages, each representing a set of linguistic objects (e.g., words, senses, synsets) and relations between them and is linguistic/terminological ontology.

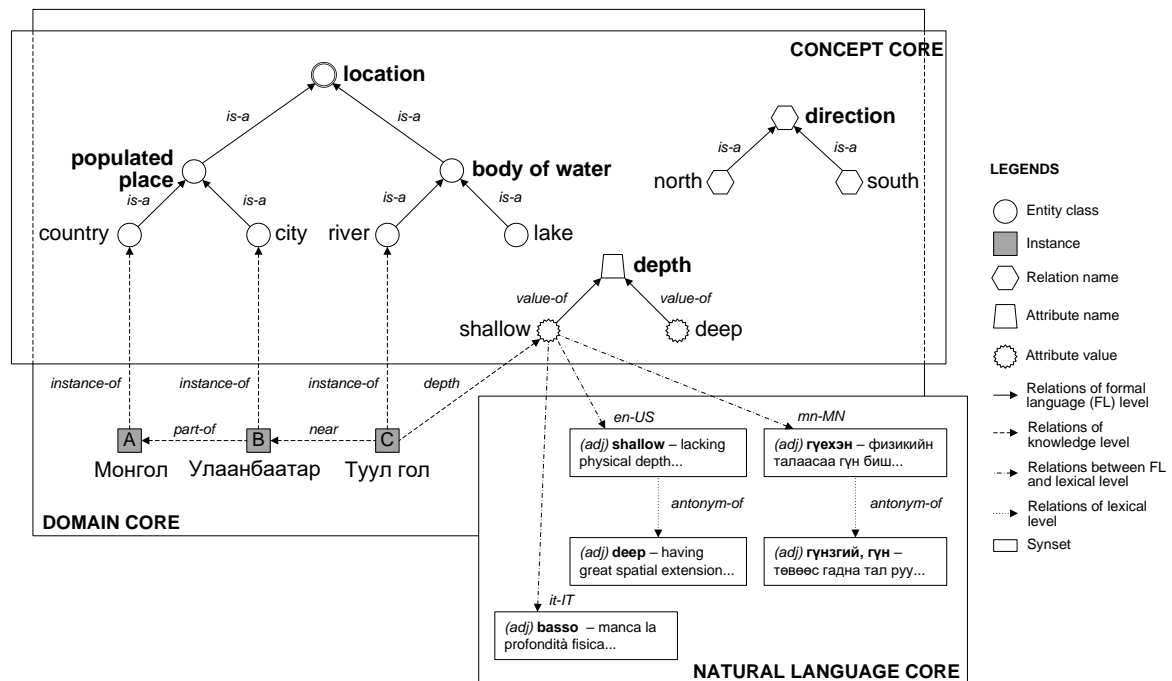


Figure 1.5 Knowledge Organization in the UKC

In order to enrich concept and natural language core with the multilingual knowledge, we need to add a new language and adapt the concept core in such integrated knowledge base. Ontology localization through translation of linguistic objects from a language to another language seems to be a promising approach which reuse of the existing concepts and their relations. When dealing with the localization issues, no lexicalization and conceptualization problems can be found in terms of multilinguality and cultural differences. In other words, there are common concepts shared among all cultures while existing specific concepts which are only used in a local community and culture. Therefore, the types of diversity across language and cultures are needed to defined and solved in a systematic and structured way.

CHAPTER 2. CROWDSOURCING

Crowdsourcing is the process of exploiting a large number of contributions made by individuals in the crowd and also taking advantage of human intelligence [71]. The crowdsourcing systems has four main challenges: recruiting contributors, human intelligent task design, combining contributions and evaluating users and contributions. This crowdsourcing technique has been used in language specific tasks, for instance generating

language resources that requires human-level quality or tasks that computers cannot do with high precision. Several researches was conducted via crowdsourcing. A workflow design of collaborative translation [52], translation tasks in library cataloguing [57] and a quality control model for non-professional translation [54], [56] referred to crowdsourcing translation challenges and evaluation of translation output. These literatures agree that iterative tasks in crowdsourcing translation including manual evaluations are inexpensive and achieve robust results with the language translation task. They also used evaluation methods including BLEU (Bilingual Evaluation Understudy), HTER (Human-mediated Translation Edit Rate), WER (Word Error Rate) and some statistical measures like Pearson and Spearman's correlations. To our knowledge, so far there is no reports on using Fleiss' kappa and Krippendorff's Alpha statistical metrics which can be used to combine users' contributions and meet our requirements of validation of ontology label translation.

CHAPTER 3. ONTOLOGY LOCALIZATION VIA CROWDSOURCING

We define a methodology for ontology localization via expert sourcing which recruits 5 types of linguistic and domain experts working with 16 elements being localized and 6 types of diversity in concept, word sense and synset level. The macro steps of the methodology is shown in Fig. 3.6. Local Knowledge Core (LKC) contains two vocabularies: source and target language and is the database where the localization activity being run and synchronized elements with the central hub UKC.

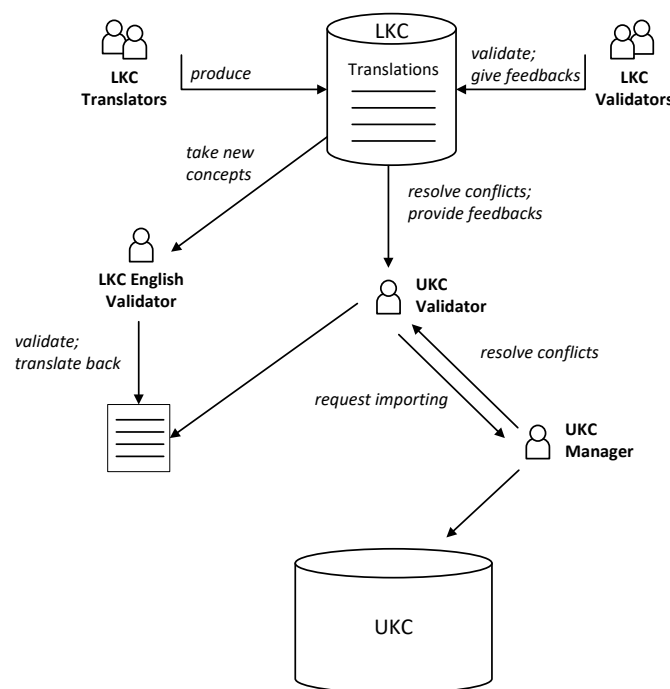


Figure 3.6 Translation phases of LKC

In the result, we have successfully translated 95.7% of the concepts of the space ontology in the UKC into Mongolian and the remaining 4.3% (42 concepts) were identified as lexical gaps. The space ontology consists of 17 facets, 985 concepts and 8.5 million entities. We also identified 99 new concepts which can be lexical gaps in the source language, in this case English. Finally, the Mongolian space ontology is produced with 1042 concepts in total.

We define a methodology for synset translation via crowdsourcing in order to reduce cost and time required by expert sourcing. This two-phase workflow is shown in Fig. 3.19.

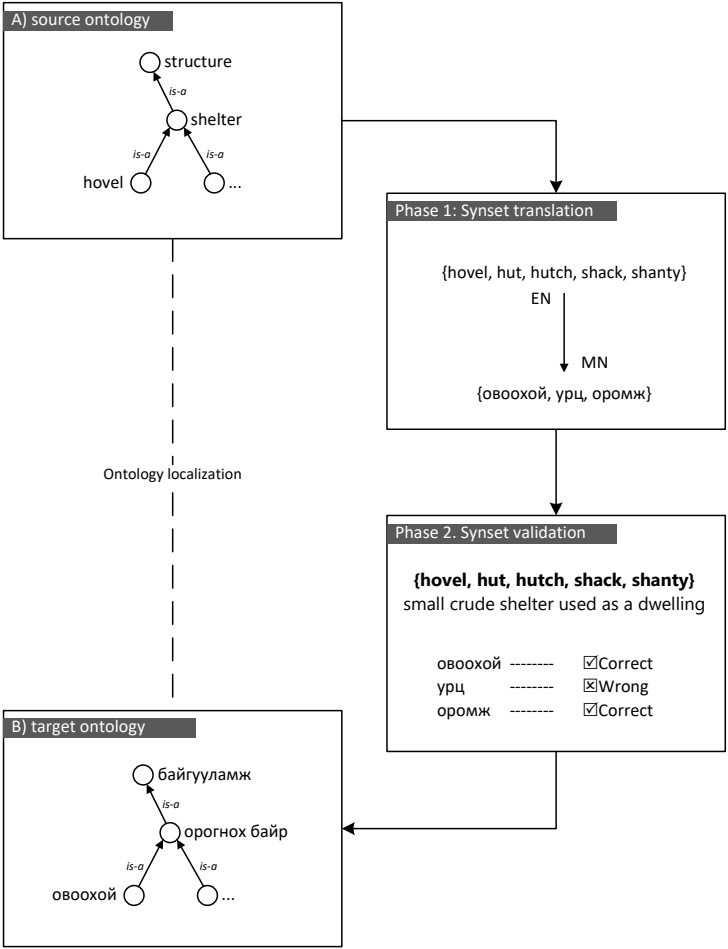


Figure 3.19 Crowdsourcing workflow of the synset localization

In phase 1, we ask 5 different contributors (web users) to provide words of synset in the target language, in this case Mongolian, which is equivalent to a given synset in the source language, in this case English or to mark the given synset as a lexical gap. The words can be varying in both synsets that means there is an added word in the target synset without direct translation of the source. This task will be asked by multiple users and they can produce many words for the target synset. Of course, user can skip a task which might be difficult to translate or unknown for him or her.

In phase 2, we again ask users to validate all distinct words in the synset by annotating the words into three categories: Correct, Wrong and Unknown. In this qualitative evaluation, we use the statistical metrics Fleiss' kappa (2.7) and Krippendorff's alpha (2.12) to compute inter-rater agreement per synset. If these correlation coefficient is higher (close to 1), the validators correctly classify the synset words which means they can separate the correct and wrong words.

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (2.7)$$

$$\alpha = 1 - \frac{D_o}{D_e} \quad (2.12)$$

We conducted our experiment in the CrowdCrafting²⁷ platform which is a web-based service that invites volunteers to contribute to crowdsourcing projects. In phase 1, 77 web users were asked 947 synsets to translate into the target language and in phase 2, 75 web users were asked to validate the results of the phase 1. In the total, all contributors have done 9,490 tasks and produced 6,442 words for the 947 synsets.

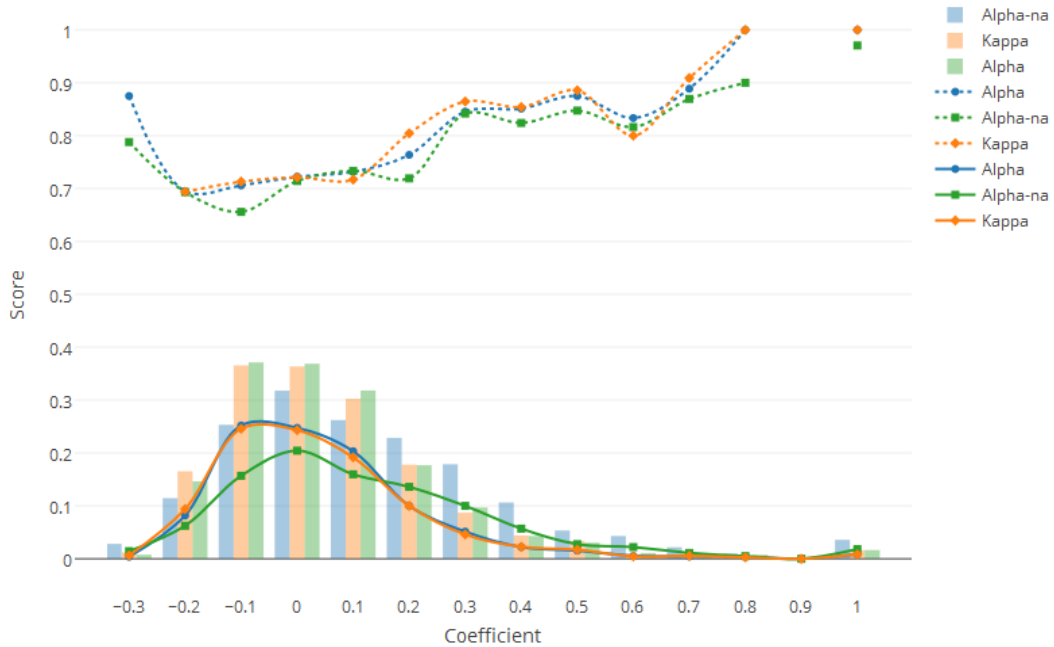


Figure 3.24 Precision, recall, F1-score of different subset of synsets

In order to evaluate contributions from the crowd, we had compiled a gold standard for the space ontology in Mongolian which is the result of expert sourcing methodology. Then, we

²⁷ <http://crowdcrafting.org>

compared this gold standard against a number of different subset and each consists of synsets whose correlation coefficient is between specified range of values (Fig. 3.24). For a synset we select words by majority votes of the correct categorization. Alpha/na/ is the correlation coefficient when the unknown category is as missing data and not included in calculations. Fig. 3.25 shows precision, recall and f1-score of different subsets whose agreement coefficient is greater than a certain value against the gold standard.

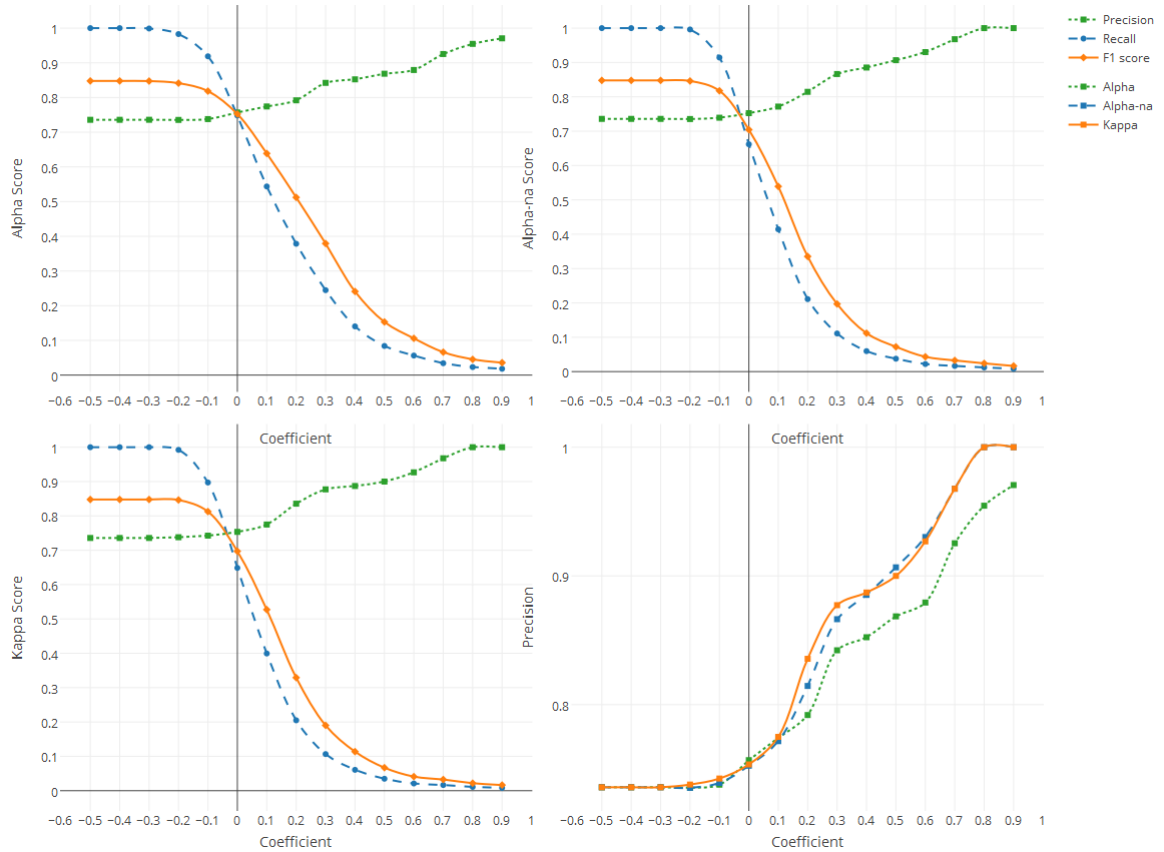


Figure 3.25 Precision, recall and f1-score (synset agreement greater than coefficient)

If the synsets whose correlation coefficient is greater than -0.1, precision increases and 0.738 for regular Krippendorff's alpha, 0.739 is for alpha with missing data and 0.743 is for Fleiss' kappa. Finally, we combine these the methodologies and first, crowd could contribute in synset translation with precision of 0.74, second the experts work on the contributions of crowd and refine the results. The comparison of this methodology against other ontology localization methodologies is shown in Table 3.15. Our methodology supports all languages as well as low-resourced language and does not require any resources in target language and NLP tools. In addition, this proposed methodology works on 15 elements of ontology and can capture diversity at language and concept level. The precision was improved to 0.74.

Table 3.15 Comparison of ontology localization methodologies

N	Characteristics / Methodology	UPM (2009) [7], [10]	NUIG (2013) [17]	BU (2016) [22]	NUM-UNITN (2013) [29], [32]
1.	Main method	Manual / Semi-automatic	Automatic	Crowdsourcing	Crowd and expert sourcing
2.	Target language resource	EuroWordNet, Wiktionary	Europarl parallel corpus	Not required	Not required
3.	Low-resourced language support	Yes	No	Yes	Yes
4.	Experts required	Yes	No	No	Yes
5.	NLP tools	Word Sense Disambiguation, Word Sense Discovery	Semantic Similarity	M-ATOLL	Not required
6.	Machine translation service	Google Translate, Babelfish, FreeTranslation	Moses Toolkit	Not required	Not required
7.	Number of elements to be localized	2 (lexical entry, lexicalization)	1 (ontology label)	1 (ontology label)	15 (synset, sense, lemma etc.)
8.	Handling diversity	Language level	No	No	Language and concept level
9.	New element creation	No	No	No	Yes
10.	Evaluation method	Gold standard	Gold standard	Gold standard	Gold standard
11.	Precision	0.72	-	>0.70	> 0.74

CONCLUSION

In this work, we proposed a methodology for generating multilingual ontologies through translation from one language into another language. While translating the ontologies, we identified 6 diversity features and generated a space ontology in Mongolian that is as high quality as the original one in English and has 1436 words and 1042 concepts including 99 new concepts from the local language and culture.

In order to reduce experts' contributions, we presented two-phase workflow for crowdsourcing synset words through translations task that we conducted an experiment and we evaluated the effectiveness of the methodology by comparing with the gold standard. We tested two statistical metrics Fleiss' kappa and Krippendorff's alpha and achieved precision of 0.74 which is higher than other methodologies. We also found out that precision increases when inter-rater agreement coefficients' value for a synset is greater than -0.1.