

ЦАХИМ ХЭЛШИНЖЛЭЛИЙН ТУХАЙ¹

Э. Мөнх-Учрал²

1. Цахим хэлшинжлэл гэж юу вэ?
2. ЦХШ-ийн түүхийн тойм
3. ЦХШ-ийн салбарууд
4. ЦХШ-д ХШ холбогдох нь

1. Цахим хэлшинжлэл (Computational Linguistics) гэж юу вэ?

Цахим тооцоолуур (computer)-ын хүч, эрчтэй хурдыг дагалдан олон салбар шинжлэх ухаан шинээр бий болон хөгжиж, үр дүн нь хэрэглээнд зогсолтгүй нэвтэрч, орон зай, цаг хугацааг эрхшээн дэлхий даяар тарж, өөрийн гэсэн ертөнцийг бүтээжээ. Харьцангуй шинэд тооцогдох цахим хэлшинжлэл хэмээх ухааныг тодорхойлохдоо:

“Цахим хэлшинжлэл бол тооцооллын үүднээс байгалийн хэлийг ойлгох, боловсруулах, бүтээж гаргах шинжлэх ухааны судалгаа бөгөөд хэлшинжлэлийн янз бүрийн онцгой үзэгдлийг тооцооллын загварт оруулах зорилготой, олон шинжлэх ухааны уулзвар дээр байдаг, хиймэл оюун ухааны салбар болсон ухаан юм” [ACL; Martin Volk 2004; Frank Richter, 2005/2006].

Хамгийн товчоор, *“Цахим хэлшинжлэл бол хүний хэлийг ойлгох, бүтээх цахим тооцоолуурын системийн судалгаа юм”* [Ralph Grishman, 1986].

Энд дурдуштай зүйл бол ЦХШ-д байгалийн хэл³ болон цахим тооцоолуурын харилцан ажиллагаатай холбоотой *“Байгалийн хэлний боловсруулалт/БХБ”* (Natural Language Processing/NLP) хэмээх чухал салбар бий. Энэ нь хүний хэлийг ойлгох, бүтээх талаараа ЦХШ-тэй нийтлэг байж, олон талаараа давхацдаг. Хэдий тийм боловч ЦХШ нь цахим тооцоолуур, хэл хоёрын хооронд харьцаа тогтоохын дээр хүний хийдгийг загварчилдгаараа өргөн хүрээтэй, БХБ нь илүү хэрэглээний талд байдгаараа ялгарна. ЦХШ-ийг хоёр талаас нь авч үзэж болно [Shuly Wintner, 2004]. Үүнд:

1. Хэлшинжлэлд зориулж цахим тооцоолуураас боловсруулсан арга барил, хэрэглээ.
2. Цахимд зориулж хэлшинжлэлийн үүднээс гаргасан арга барил, хэрэглээ.

Эхнийх нь хэлний тухайлсан үзэгдлийг тооцооллын үүднээс тайлбарлахад чиглэсэн судалгаа, дараах нь цахим тооцоолуурыг яриа ойлгодог, хэлийг хооронд нь

¹ Computational Linguistics/CL гэдгийг “Цахим хэлшинжлэл” гэх саналтай байна. Учир нь, Чой. Лувсанжав гуай “цахим” гэсэн үгийг нэр томьёоны талбарт оруулж, *цахим тооцоолуур* (computer), *цахим захиа* (e-mail) зэргээр хэвшүүлж буй [Шагдарсүрэн 2007]. Тус өгүүлэлд нэр томьёог монголчлохыг хичээсэнд маань засаж залруулах зүйл магад бийг болгооно биз ээ.

² МУИС-ийн Монгол хэл, соёлын сургууль.

³ *Байгалийн хэл* (Natural Language) нь *хиймэл оюун ухаан* (Artificial Intelligence) бүхий цахимын ертөнцөд зориулан загварчилсан “C++”, “Java” зэрэг хэлнээс ялгах зорилгоор хэрэглэдэг, хүний хэл (Human Language)-ийг илэрхийлсэн нэр томьёо. ӨМӨЗО-нд “Байгаж хэл” гэж оноосон.

орчуулдаг болгох зэрэгт анхаарч машиныг хэлний хувьд хүнтэй адил түвшинд харилцдаг болгоход чиглэсэн илүү технологийн шинжтэй судалгаа юм.

ЦХШ-ийн хамгийн чухал онцлог бол “**цогц нэгдмэл байдал**” юм. Учир нь:

- *Нэгд*, цахимын ухаан, хэлшинжлэл, математик, логик, сэтгэл судлал гэх мэт олон шинжлэх ухааны уламжлалт судалгаанд тулгуурлаж хөрс сууриа болгодог.
- *Хоёрт*, эхний онцлогтой холбоотойгоор хэлшинжлэлч, цахимын эрдэмтэн, хиймэл оюун ухааны мэргэжилтэн, математикч, логикч, сэтгэхүйн ухаанч, хүн судлаач гэх мэт эрдэмтдийн багийн хамтын ажиллагааны нэгдэлд хөгждөг,
- *Гуравт*, загварчлал нь хэлшинжлэлийн ямар нэг түвшнээр хязгаарлагдахгүй, хэлний олон түвшний цахимд зориулсан цогц мэдлэг бүтээсний дүнд бодтой үр дүн нь гардаг.



ЦХШ-ийн гол хандлага нь *дүрэмд суурилсан тогтолцоо* (Rule-Based Systems), *өгөгдөлд суурилсан тогтолцоо* (Data-Driven Systems) юм [Frank Richter, 2005/2006].

Дүрэмд суурилсан тогтолцоо нь хэлшинжлэлийн мэдлэгийг тодорхой илэрхийлсэн кодчилолтой, зориуд бэлдсэн жишээ бүхий хязгаарлагдмал бөгөөд шалгагдсан өгөгдөл, гараар оруулсан хэлзүйн дүрэмд тулгуурлана. Ингэснээр турших, алдаа илрүүлэхэд хялбар байдаг. Хүний нөр хичээл зүтгэлээр бүтээнэ. Оруулж буй өгөгдлийн хүрээ өргөсөх тусам бэрхшээлтэй болохын дээр, хэрэв өгөгдөл буруу байвал хэрэглээний найдвартай, хүртээмжтэй байдал ямар нэг хэмжээгээр алдагдана.

Өгөгдөлд суурилсан тогтолцоо нь хэлшинжлэлийн мэдлэгийг далд илэрхийлсэн кодчилолтой, статистик аргад тулгуурладаг тул хүний оролцоо харьцангуй бага. Өмнөх тогтолцоог бодвол зөв бус өгөгдөлтэй зохицох чадвар илүү уян хатан. Харин өгөгдлийн үлэмж сурвалж шаардаж, үр дүн нь түүнээсээ шууд хамаардаг.

2. Цахим хэлшинжлэлийн түүхийн тойм.

Хэлийг боловсруулах, хэлний хэрэглээг хангахад цахим тооцоолуур ашиглах санаа төрсөн шалтгаан нь хүний мэдлэгийг ашиглахад цаг, хөлс, оюуны болоод төлөвлөгөөт ажлын боломжоос хараат буюу товчоор их үнэтэй байдаг бол цахим тооцоолуур арифметик үйлдлийг хүнээс хурдан, оновчтой гүйцэтгэж чадах нь нотлогдсоноор илүү бага зардал зарцуулж болно хэмээн төсөөлсөнд буй.

Гэтэл хамгийн их бодолцох ёстой зүйл бол хэлний мөн чанар юм. Хүн төрөлхтөнд үг яриа, бичвэрийг хурдан түргэн ухаж ойлгох, хэлмэрчлэх ер бусын авьяас чадвар заяажээ. Хэдий энгийн санагдаж болох ч энэ далд нууцлаг үйлийн эс ширхэг, бүтэц бүрэлдэхүүн нь машинд ойлгуулахад тийм авсаар зүйл биш ээ. Учир нь, хүний ухамсрын тогтолцоонд цахим тооцоолууртай харьцуулбал, илүү нухацтай хандах ёстой чухал үйл явцууд бий. Юуны өмнө, энэ үйлдэл маш хурдан болдог. Зарим үед хэдэн үг, өгүүлбэр уншиж, сонссоноор гол санааг барьж, орхигдсоныг нөхөж, сэдвийг бүхэлд нь

ойлгож болно. Өөр нэг чухал зүйл бол хэлэнд “байнгын тогтвортой” бүтэц байхгүй, цагийн эрхээр өөрчлөгдөж байдаг. Хүн төрөлхтөн хэлний тэрхүү хөдлөнги чанартай маш сайн харьцдаг. Эд бол цахимын ертөнцтэй нэг жишигт оруулсан хэрэгслээр дамжиж ажиллахаар хашигдсан бидэнд тулгардаг гол бэрхшээл юм. Жорж Миллер энэ тухай “Ихэнх хүмүүс хүний хэл үнэндээ ямар бүрхэг бүдэг, утгыг янз бүрийн аргаар илэрхийлдгийг төдий л сайн мэддэггүй тул цахим тооцоолуур хэлний харилцааг ойлгож чадахгүй байгааг хараад урам хугардаг. ...Цахим тооцоолуур үүнд буруугүй бөгөөд хэл өөрөө үүнийх нь шалтгаан болдог” [G. Miller, 2001:26] хэмээн оновчтой илэрхийлжээ.

Хэлийг цахимд боловсруулахын тулд түүний хөдлөнги шинжтэй холбогддог учраас энэ судалгаа эхэлсэн цагаасаа эдүгээ хүртэл олон жил зарж, түүнчлэн төгс үр дүнд хүрээгүй байсаар буй.

1940-1950-аад он:

Орчуулгад цахим тооцоолуур ашиглах санаа Вийверт 1947 онд төрж [Hutchins, 1997], 1949 онд тэмдэглэл [Locke, 1955:15-23] бичсэнээр бодтой биелэх эхлэл тавигджээ. 1954.1.8-нд “New York Times” болон бусад сонин цахим тооцоолуураар орчуулсан тухай мэдээг олон нийтийн сонорт хүргэв. Жоржтаун их сургууль, “IBM” компанийн хамтарсан “The Georgetown-IBM experiment” хэмээх төслийн дүнд 60 гаруй өгүүлбэрийг оросоос англи руу орчуулсан бөгөөд ердөө 6 хэлзүйн дүрэм, 250 үгтэй байв. Зохиогчид 3-5 жилд машин орчуулга /МО (Machine translation/MT) шийдэгдэх боломжтой гэж мэдэгджээ [Hutchins, 2005]. Тус төсөл МО байх боломжтойг харуулж, хүмүүсийн сонирхол, санхүүжилтийг татсан анхдагчийнхаа хувьд туйлын үнэ цэнтэй.

Ийнхүү 1950-аад оны үест “Цахим хэлшинжлэл” хэмээх ухаан АНУ-д үүссэн бөгөөд түүхэн үүднээс хамгийн анхны, арваад жилийн турш цахим хэлшинжлэлийн цорын ганц байсан салбар бол МО юм. Энэ тухай Вилкс “БХБ буюу ЦХШ-ийн 50 жилийн түүхэн дэх тэмдэглүүштэй нэгэн онцлог бол энэ салбарын бүхий л анхдагчид цахим тооцоолуураар гачигдаж байсан тэр л эхлэлийн үед ямар их сонирхолтой сэдэв, агуулгын нөхцөлийг бидэнд гаргаж өгч чадсан юм бэ гэдэгт оршино” [York Wilks, 2005:1] хэмээн онцлон тэмдэглэсэн билээ.

Энэ үеийн ЦХШ-д холбогдох онолын хөгжил нь хийсвэр хэлний онол, мэдээллийн онол юм.

Хийсвэр хэлний онол (Formal Language Theory): “Хийсвэр хэл бол илүү ч үгүй, дутуу ч үгүй төгсгөллөг урттай стрингүүд буюу үгийн иж юм” хэмээн тодорхойлдог. Стрин (String) нь токен (token), тэмдэг (symbol), үсэг (letter) байж болно. Хийсвэр хэл нь хийсвэр хэлзүй (formal grammar)-гээр тодорхойлогдох ба өгөгдлийн хийсвэржил болон програмчлалын хэлний өгүүлбэрзүйн яг таг тодорхойлолтод ашиглагдаж үгийн сангийн задлуур (lexical analyzer), хэлзүйн задлуур (parser) зэргийн хөгжилд гол үүрэг гүйцэтгэдэг. Онолын гол төлөөлөгчид нь Чомский⁴, Клеене⁵, Бакус⁶ нар юм. Хэлний хийсвэрлэлийн үзүүлэлтийг дараах байдлаар тодорхойлно:

⁴ **Чомский** (Avram Noam Chomsky 1928-) 1950-аад оноос Үүсгүүр хэлзүй (Generative Grammar)-н онолоо хөгжүүлж эхэлсэн нь боловсорсоор хэлшинжлэлд өөрийн байр суурийг олсон. Түүний хэл болон хэлзүйн хийсвэрлэлийн тодорхойлолт нь хэлний илэрхийллийн автомат өгүүлбэрзүйн боловсруулалтын үндсийг хангаж өгдөг. Үүсгэх хүчнийх нь нөхцөлөөр хийсвэр хэлийг ангилсан Чомскийн шатлал (Chomsky hierarchy) нь хийсвэр хэлний онолын хөгжилд чухал үүрэг гүйцэтгэсэн [Chomsky, 1956]. Энэ нь: Type-0: хязгаарлагдмал бус хэлзүй (unrestricted grammars), Type-1: хам сэдэв мэдрэх хэлзүй (context sensitive grammars) Type-2: чөлөөт хам сэдвийн хэлзүй (context free grammars) Type-3: энгийн хэлзүй (regular grammars).

1. Стрингүүд нь хийсвэр хэлзүй (formal grammar)-гээр үүсдэг байх,
2. Тогтсон илэрхийлэл (regular expression)-ээр дүрслэгдэх,
3. Тюрин машин⁷, төгсгөллөг төлөвт автоматчилал⁸ зэрэг автоматчиллаар дэмжигдэх,
4. Зарим шийдлийн үйл явц (decision procedure)-ад “тийм” гэсэн хариулт өгдөг байх.

Хэрэглээний математик болон цахим инженерчлэлийн салбар болох **мэдээллийн онол** бий болсноор БХБ, нууцлал зэрэг хэрэглээний бусад талбарыг өргөжүүлж өгчээ. Энэ онолын ойлголт нь хэлний загварын үр дүнг хэмжих зэргээр ашиглахад ач холбогдолтой. Гол төлөөлөгч нь Шенон⁹.

1960-1980 -аад он

Хэлийг барилга гэвэл уг тоосго нь юм. Нэг хэлийг нөгөө рүү орчуулахын тулд дан үгээс гадна үгзүй, өгүүлбэр зүйг багтаасан хэлзүйг цахимд зориулан судлах шаардлага гарч, улмаар өгүүлбэрзүйг ойлгуулахын тулд утгазүй, үгийн сан, цаашлаад хэлний хэрэглээний талыг таниулах хэрэгтэй болох зэргээр асуудал шил даран хөвөрч эхэлсэн. Тиймээс анхны төслийн багаас мэдэгдэж байсан шиг амар байсангүй.

Машин орчуулгын түүхэнд дуулиан тарьсан үйл явдал бол эргэлзээгүй 1966 оны “ALPAC”¹⁰ төслийн тайлан билээ. Энд “...Үгээс уг рүү орчуулах орчуулга ажиллахгүй юм байна. Хэлшинжлэлийн мэдлэг хэрэгтэй. ...Машин орчуулга ойрын ирээдүйд боломжгүй...” хэмээн дүгнэжээ [Claire Gardent, 2003:6]. Үүний үрээр АНУ дахь машин орчуулгын санхүүжилт хорь орчим жилээр зогссон аж [Hutchins, 1996, 2003]. МО зорьсондоо хүрээгүй энэ үеэс хүний хэлийг боловсруулахад анхны төсөөллөөс илүү иж бүрэн судалгаа хэрэгтэйг ойлгож авчээ.

1960-1980-аад оныг тэмдгийн (symbolic) буюу магадлалын (stochastic) загварчлалын үе гэж хэлж болно. Энэ үеийн судалгааг товч өгүүлбэл,

Тэмдгийн шинжтэй эх хэлний боловсруулалтын амжилт:

- Хэлийг боловсруулах ба сурах системийн үндэс болсон хийсвэр хэлзүйн хэрэглээ (Чомский, Харрис¹¹)

⁵ **Клеене** (Stephen Cole Kleene 1909-1994) Математикийн зөн совингийн талынхныг тэргүүлж байсан тэрбээр Клеене од, Клеене шатлал, Клеене алгебр зэрэг математикийн ойлголтыг гаргаж, цахимын онолын үндсийг тавихад хувь нэмрээ оруулсан.

⁶ **Бакус** (John Warner Backus 1924-2007) Хийсвэр хэлний өгүүлбэрзүй (formal language syntax)-г тодорхойлоход хэрэглэдэг “Backus-Naur form” (BNF)-ийг зохиосон. “BNF” нь дээд түвшний өгүүлбэрзүй (metasyntax) бөгөөд чөлөөт хам сэдвийн хэлзүй (context free grammar)-г илэрхийлэхэд хэрэглэгдэх ба үгийн сангийн болон өгүүлбэрзүйн дүрмээс бүрддэг.

⁷ **Тюрин машин** (Turing machine) 1937 онд Тюрин (Alan Turing, 1912-1954) уг онолын хэрэгслийг дүрсэлсэн. Энэ нь ямар нэг алгоритмын логикийг адилхан үзүүлэхээр өөрчлөгдөж чаддаг ба механик тооцооллын хязгаарыг ойлгоход тусалдаг.

⁸ **Төгсгөллөг төлөвт автоматчилал** (Finite state automaton) нь төлвийн төгсгөллөг тоогоор зохиосон, төлвийн хооронд шилжүүлэг ба үйлдэл бүхий загвар. Шилжүүлэг нь төлвийн өөрчлөлтийг зааж, шилжүүлэх боломжтой нөхцөлөөр дүрслэгдэнэ. Үйлдэл нь орох (entry action), гарах (exit action), оруулах (input action), шилжүүлэх (transaction action) гэсэн хэлбэртэй.

⁹ **Шенон** (Claude Elwood Shannon 1916-1994) “Мэдээллийн онолын эцэг” хэмээн өргөмжлөгддөг түүний ЦХШ-д оруулсан хувь нэмэр нь цагаан толгойн 27 дахь үсэг болох “хоосон зай” (space)-н боловсруулалт бичгийн хэлэнд хэдийгээр тодорхой бус зүйл байдаг ч тоон тодорхойлолтод ордог “илэрхий” холбоос юм гэдгийг нотолж бичсэн өгүүллээс эхэлжээ [Shannon 1951].

¹⁰ **Пийрс** (John Pierce 1910-2002)-ийн тэргүүлсэн “ALPAC” (Automatic Language Processing Advisory Committee) хороог 1964 онд АНУ-ын ЗГ-аас машин орчуулгыг хөгжүүлэхээр байгуулсан.

¹¹ **Харрис** (Zellig Harris 1909-1992) Бүтэц хэлшинжлэл болон харилцан ярианы задлан шинжилгээний талбар дахь ажлууд, хэлний хувьсанги бүтцийн нээлтээрээ алдаршсан. Сапир (Edward Sapir, 1884-1939) түүнийг өөрийнхөө оюуны залгамжлагч хэмээн үзэх болсон. Түүнийг

- Утгын болон өгүүлбэрийн гаргалгааг тодорхойлох учир шалтгаанд суурилсан програмчлал ба учир шалтгааны хэрэглээ (Каплан, Кей¹²)
- Хэлийг ойлгож, үүсгэдэг анхны систем (Мински¹³, Шанк¹⁴, Виноград¹⁵, Колмераур¹⁶)
- Харилцан яриа боловсруулалт (Гросз¹⁷, Сиднер¹⁸)

Магадлалын загварчлалын хувьд:

- Яриа танихад зориулсан магадлалын аргууд (Бледсое¹⁹, Броунин²⁰, Желинек²¹)

Тэмдгийн шинжтэй боловсруулалтад тулгуурлан хэлшинжлэлийн буюу учир шалтгааны үндэслэлтэй, тооцоололд суурилсан загварууд гарав. Үүнд, *мөчирлөсөн хэлзүй*²², *"Montague PTQ"*²³, *үгийн сангийн үүргийн хэлзүй*²⁴, *ерөнхийлсөн хэллэгийн*

Блүүмфийлд (Leonard Bloomfield, 1887-1949)-ийн ХШ-ийн тодорхойлолт дүрслэлийн санаануудыг дээд түвшинд нь хүргэсэн гэж үздэг. Үүнд, авиабар ба бүтээврийн тархаж байрлах шинж чанар (distributional properties)-т үндэслэн хэлний нэгжийн үйлдлийг таних шинжилгээг дурдаж болно. *Бүтэц хэлшинжлэл дэх аргууд* [Harris, 1951] хэмээх алдаршуулсан бүтээл нь түүний олон жилийн ажлынх нь эцсийн томъёолол, боловсруулалт юм. Тэр бээр үр дүнг нотлох тодорхой аргууд нь өгөгдлөөс хангалттай хэмжээнд үүснэ гэж үзсэн.

¹² **Каплан** (Ronald M. Kaplan), **Кэй** (Martin Kay) нар төгсгөлөг төлөвт (finite state) авиазүй, үгзүйн дүрслэлийн хэрэглээний доорх математик, хэлшинжлэл, тооцооллын ойлголтыг хөгжүүлсэн. **Кей** "МО нь хүний орчуулах үйл явцыг бүрэн бий болгож чадна" гэж үздэг.

¹³ **Минский** (Marvin Lee Minsky, 1927-) Гүүглийн лавлахад салбарынхаа бүх цаг үеийн 6 алдартны нэг гэж орсон байдаг. Паперт (Seymour Papert)-ын хамтаар анхны "Logo turtle"-ийг хөгжүүлсэн.

¹⁴ **Шанк** (Roger Schank 1946-) 1970, 80-аад онд ХОУ, сэтгэл судлалын нөлөө бүхий хөгжүүлэгчдийн нэг байсан түүний гол шинэчлэл нь *тохиолдолд суурилсан бодлого* (case based reasoning), *хөдлөнгү санах ой* (dynamic memory)-н тухай ойлголт болно.

¹⁵ **Виноград** (Terry Allan Winograd 1946-) "SHRDLU" бол түүний докторын зэрэг горилсон зохиол.

¹⁶ **Колмераур** (Alain Colmerauer, 1941-) Логик програмчлалын хэл "Prolog"-ийг зохиосон бөгөөд "Constraint logic programming"-ын талбарыг үндэслэгчдийн нэг.

¹⁷ **Гросз** (Barbara J. Grosz) Тэр бээр цахимын харилцан ярианы системийг хөгжүүлж, ярилцлагын тооцооллын загварчлалын талбарыг бий болгосон. Ярилцлагын ач холбогдол нь ярьж буй хүний санаа зорилго, анхаарлын байдал, хэлний хэлбэрийн харилцан үйлчлэлээс хэрхэн хамаарч буйг тусгайлан үздэг *ярилцлагын бүтцийн онол* (theory of discourse structure)-ыг хөгжүүлж ирсэн. [Barbara Grosz, Candace L. Sidner, 1986]

¹⁸ **Сиднер** (Candace L. Sidner) Хүний харилцаа, хамтын ажиллагаа, түүний агент, робот, дүр төрхийн хэрэглээний програмуудын талаар олон жил судалсан. <http://mysite.verizon.net/sidner/>

¹⁹ **Бледсое** (Woodrow Bledsoe, 1921-1995) Автоматчилсан теоремын баталгаа, хээ таних салбарын анхдагчдын нэг. Бледсое, Броунин нарын ажлын үр дүн(1959)-нд хэвлэмэл болон гар бичмэл захиаг таних зорилготой "N-tuple" арга гарсан.

²⁰ **Броунин** (Iben Browning, 1918-1991) ХОУ, био-инженерчлэл, цаг уур судлал, амьтан судлал, философи, бизнес зэрэг маш олон салбарт ажиллаж байсан.

²¹ **Желинек** (Frederick Jelinek, 1932-) Мэдээллийн онол, автоматаар яриа таних, БХБ-ын эрдэмтэн бөгөөд тоо бүртгэлийн загварчлал хэрэглэхийг санаачилсан.

²² **Мөчирлөсөн хэлзүй** (Tree Adjoining Grammar) Энэхүү хэлзүйн хийсвэрлэлийг 1969 онд Аравинд Жоши (Aravind Krishana Joshi, 1929-) тодорхойлсон. Зарим талаар чөлөөт хам сэдвийн хэлзүйтэй төстэй боловч боловсрууллын энгийн нэгж нь модоор дүрслэгддэг.

²³ **Montague PTQ** (The Proper Treatment of Quantification) Хийсвэр логикт суурилсан, утгын боловсруулалтад чиглэсэн хандлага болох хийсвэр утгазүй (Formal semantics)-н судалгаа. Монтэг (Richard Montague, 1930-1971) байгалийн хэл (NL) ба хийсвэр хэл (FL) (програмчлалын хэл (PL) шиг)-ийг нэг замаар боловсруулж болно гээд "...Миний бодлоор байгалийн хэл ба хиймэл хэл (artificial language)-ний хооронд онолын онц ялгаа байхгүй. Чухам тиймээс энэ хоёр хэлний өгүүлбэр зүй, утгазүйг байгалийн болоод математикийн нарийн тодорхойлсон нэг онолын хүрээнд ойлгож ухамсарлах боломжтой гэж үзэж байна" хэмээжээ. [Montague 1970] Түүний 3 өгүүлэл *Монтагийн хэлзүй* хэмээн алдаршсан. [Montague, 1970a, 1970b, 1973]

бүтцийн хэлзүй²⁵-н шууд үргэлжлэл болох толгой удирдлагат хэллэгийн бүтцийн хэлзүй²⁶, оператор хэлзүй²⁷ зэргийг дурдаж болно. Эдгээрт суурилсан цөөн хэдэн хэрэглээний програм гарсан ч хангалттай зөв үр дүн үзүүлэхгүй байв.

²⁴ **Үгийн сангийн үүргийн хэлзүй** (Lexical Functional Grammar) 1970-аад онд Бреснан (Joan Bresnan 1945-), Каплан (Ronald Kaplan) нар эхлүүлсэн үүсгүүр хэлзүй-н нэг төрөл бөгөөд өгүүлбэрийг үг зүй, утгазүйн холбооных нь хамтаар анхаарч хэл бол олон хэмжигдэхүүнт бүтэцтэй, хэмжигдэхүүн бүр нь өөрийн гэсэн дүрэм, ойлголт, хэлбэр бүхий онцгой бүтцээр илэрхийлэгдэнэ хэмээдэг. Бүтцийн бүрэлдэхүүний түвшин бүрийг хоорондын дарааллаар нь дээш зураглан багтаадгаараа онцлог. 1. Хэлзүйн үүргийн дүрслэл (the representation of grammatical functions/f-structure), 2. Өгүүлбэрзүйн бүрэлдэхүүний бүтэц (the structure of syntactic constituents/c-structure) гэсэн бүтцийг авч үзнэ. Жишээ нь, "The old woman eats the falafel" гэсэн өгүүлбэр байхад "c-structure" нь өгүүлбэр нэрийн болон үйлийн хэлц (NP&VP) гэсэн 2 үндсэн хэсгээс бүтнэ, үйлийн хэлц нь үйл үг ба нэрийн хэлц -ээс бүтнэ зэргээр бүтцийг үг хүртэл гаргана. "f-structure" нь тоо, цаг (number, tense) гэх мэт онцлог, өгүүлэгдэхүүн, өгүүлэхүүн, тусгагдахуун (subject, predicate, object) зэрэг нэгжийн шинжийг боловсруулна. Хожим маргааны (argument structure), утгын (semantic structure), мэдээллийн (information structure), үгзүйн (morphological structure), авиазүйн (phonological structure) зэрэг бүтэц бий болсон. УСҮХЗ-г баримтлагчид хэлний түгээмэл загвар байх боломжтой гэж үздэг. Судалгааны ололт нь чанд хийсвэрлэлтэй хэлзүйн нарийн загварыг бий болгодогт оршино. Иймээс "Lekta", "TranSphere" зэрэг МО-ын хэрэгслийн онолын суурь болдог. "Lekta" бол Жулиетта (Julietta Research Group)-аас гаргасан 2 хэлний МО-ын хэрэгсэл. "TranSphere" нь "AppTek" компанийн дээд төвшний хэлшинжлэлийн програмчлал ашиглахаар загварчилж тоногдосон, англи ба бусад хэл гэсэн 2 чиглэлт (bidirectional) орчуулгын програм.

²⁵ **Ерөнхийлсөн хэллэгийн бүтцийн хэлзүй** (Generalized phrase structure grammar /GPSG) бол 1970-аад оны сүүлээр Газдар (Gerald Gazdar, 1950-) анхлан хөгжүүлсэн, өгүүлбэрзүй, утга зүйн дүрслэлд зориулсан суурь судалгаа юм [Gerald Gazdar et al, 1985]. Өгүүлбэрзүйг чөлөөт хам сэдвийн хэлзүйгээр дүрсэлж болно гэдгийг харуулснаараа чухал ач холбогдолтой. Газдар болон бусад олон өгүүлбэрзүйч хэлийг чөлөөт хам сэдвийн хэлзүйгээр хангалттай дүрсэлж чадахгүй гэж маргадаг. www.informatics.sussex.ac.uk/research/groups/nlp/gazdar/briscoe/gpsg.html

²⁶ **Толгой удирдлагат хэллэгийн бүтцийн хэлзүй** (Head-Driven Phrase Structure Grammar/HPSG) Поллард (Carl Pollard 1947-), Сар (Ivan Sag, 1949-) нар 1985 онд чөлөөт хам сэдвийн хэлзүй (CFG)-н өргөтгөл болох дээд хэмжээгээр үгийн санжсан, үүсмэл бус, үүсгүүр хэлзүйн тус онолыг хөгжүүлсэн. Энэ нь хэлзүйн дүрэм болон тэр болгон хэлзүйд хамааруулж үздэггүй үгийн сангийн нэгж (lexicon entries) -ийг багтаана. Соссюр (Ferdinand de Saussure, 1857-1913)-ийн тэмдгийн үзэл ойлголт (notion of the sign)-ыг ашиглах ба жигд хийсвэрлэлийг хэрэглэж, модулийн аргаар зохион байгуулснаараа анхаарал татдаг. Хийсвэрлэл нь үгийн санжилд тулгуурлана. Ингэснээр үгийн сан бол толгой үгийн жагсаалтаас өргөн утга агуулгатай, баялаг бүтэцтэй байна. Толгой үг бүр хэлбэрээр тэмдэглэгдэх ба хэлбэр нь шатлалтай. Үг болоод хэллэг нь тэмдгийн хоёр өөр дэд хэлбэр (subtype)-тэй. Үг нь "PHON" буюу авианы, "SYNSEM" буюу өгүүлбэрзүй, утгазүйн хоёр онцлог шинж (feature)-тэй. Энэ онцлог шинж нь дэд онцлог шинж (subfeatures) болон задарна гэж үздэг. [Pollard, Sag, 1994] Поллард мөн дээд эрэмбийн хэлзүй (Higher-order grammar)-г зохиогч ба "Convergent grammar/CVG" дээр ажиллаж буй аж.

²⁷ **Оператор хэлзүй** нь хэл мэдээллийг хэрхэн дамжуулдгийг тайлбарласан математикийн онол юм. Оүний хэл бүхэн бол өөрийн гэсэн зохион байгуулалттай систем, ямар нэг үгийн өгүүлбэрзүй, утгазүйн шинж чанар нь бусад үгтэй холбогдох байдлаас шууд бий болно гэж үздэг тул дүрмийг тодорхойлоход дээд түвшний хэл (metalanguage) шаардлагагүй гэдэг. Энд 3 шаардлага бий: 1. Хамаарал (Dependency): Үг нь илэрхийлэл болж хэлбэржихдээ бусад үгээс хамаарна. 2. Бололцоо (Likelihood): Үгийн зарим хослол, тэдгээрийн хамаарал нь бусдынхтайгаа ойролцоо байна. 3. Бууруулал (Reduction): Ойролцоо байх боломж ихтэй хослол дахь үгс нь товчлогдож, зарим тохиолдолд гээгдэж ч болно. Эдгээр нь хамтдаа хэлний мэдээллийн онол (Theory of Language Information) болох ба хамаарал нь батлах бүтцийг байгуулж, бололцоо нь тодорхой утгыг бүтээж, бууруулал нь харилцааны цомхон хэлбэрийг олгодог. [Harris, 1988, 1989, 1991]

1960-80 –аад оны амжилтууд:

1966 онд Вайценбаумын бүтээсэн харилцан яриа дуурайдаг анхны хөөрөлдөгч бот (chatbot) “ELIZA” [Weizenbaum, 1966] бол БХБ-ын эрт үеийн нэгэн ололт юм. Хэрэглэгч өгүүлбэр бичихэд, “ELIZA” оруулбар (input)-ыг уншаад, түлхүүр үгийг шалган олж, өгүүлбэрийг түлхүүр үгтэй холбоотой дүрмээр хувиргаж задлан шинжлээд, оруулбарыг сэргээн хувиргаж хариу бичвэрийг үүсгэнэ.

Энгийн харилцан ярианы жишээ:

- Men are all alike.
- IN WHAT WAY?
- They're always bugging us about something or other.
- CAN YOU THINK OF A SPECIFIC EXAMPLE?
- Well, my boyfriend made me come here.

“ELIZA” сэтгэл мэдрэлийн эмчийн яриагаар хязгаарлагддаг, хэлшинжлэлийн мэдлэгийг багтаагаагүй. Гэвч эрт үеийн цахим ертөнц, тэр дундаа тоглоомын хувьд интерфэйс зохиомжийн нэмэлт төрлүүдээрээ хүчтэй нөлөөлсөн ба түүнд түшиглэсэн олон програм олон хэлээр (Abuse-1980, The Prisoner-1980, Zebal-1993) гарч, зохиогчийг цахимын хязгаарыг тайлбарласан алдарт бүтээлээ [Weizenbaum, 1976] туурвих үндэс болсон.

“TAUM Meteo”

“TAUM” багийнхан²⁸ “TAUM-73”, “TAUM METEO” зэрэг эртний машин орчуулгыг Колмераурын бүтээсэн програмчлалын хэл “Q-Systems” [Colmerauer, 1969] ийг ашигласан. Эдгээр нь хэлшинжлэлийн задлан шинжилгээний дагуу автомат орчуулга хийх анхны оролдлого байв. 1970 онд гарсан “TAUM METEO” нь англиас франц хэл рүү цаг агаарын мэдээ орчуулдаг. Үүний нэг онцлог бол үр дүн нь ямар нэгэн хяналт засваргүйгээр олон нийтэд хүрдэг байсан.

“SYSTRAN”²⁹:

1968 онд оросоос англи руу, 1973 онд англиас орос руу, 1975 онд англи сурвалжийг орчуулдаг байсан бол өнөөдөр араб, хятад, герман, франц, япон зэрэг 20 гаруй хэлний орчуулга хийдэг болон хөгжжээ. Боловсруулалтын түвшин нь:

1. *Үгийн шинжилгээ*: Үүнд толь бичгээс хайх болон үгзүйн шинжилгээ хийх үйлдэл хамаардаг. Үйлдлийг толь бичиг, чиглүүлэгчээр дамжин гүйцэтгэнэ. Язгуурын (Stem dictionary), *давтамж ихтэй үгийн* (High-frequency dictionary), давтамж багатай үгийн (Low-frequency dictionary), үндсэн (Main dictionary) гэсэн толь бичгүүдийг ашиглана.
2. *Өгүүлбэр зүйн шинжилгээ*: Өгүүлбэрийг задлах үйлдэл хийнэ.
3. *Нэгтгэн дүгнэх*: Задалсан өгүүлбэрийн орчуулгыг хийнэ.

“LUNAR”

1973 онд гарсан Аполлогийнхны авчирсан чулууны геологийн шинжилгээний талаарх асуултад хариулах систем. Каплан Сарны систем (Lunar system)-д зориулж англи хэлний анхны томоохон хэмжээний “ATN”³⁰ хэлзүйн дүрмийг бичсэн.

²⁸ “TAUM” (Traduction Automatique a l'Universit% de Montr%al) 1965 онд Канадын Монреал их сургуульд байгуулагдаж, Жон Шандью (John Chandioux)-ийн ахалж байсан судалгааны баг.

²⁹ “SYSTRAN” - Петер Тома (Peter Toma) энэ компанийг 1968 онд байгуулж тус орчуулгын системийг хөгжүүлж, арилжаанд нийлүүлсэн. [Toma, 2000] <http://www.systransoft.be/>

³⁰ “ATN” (The Augmented Transition Network) Хийсвэр хэлний үйлдлийн тодорхойлолтод хэрэглэдэг графын онолын бүтцийн нэг хэлбэр. Төгсгөллөг төлөвт машин (Finite state machine/FSM)-ыг өгүүлбэрийн задлалд ашиглах санаанаас бий болсон.

SHRDLU³¹

Виноград 1968-1970 онд хүний хэлийг ойлгодог програм “SHRDLU”-г хөгжүүлсэн нь өгүүлбэр зүйн задлал, утгын шалтгааныг ашигласан хязгаарлагдмал талбарт хэлийг ойлгоно, англи хэлний дүрэм ба хэлзүйн задлуур (parser)-тай, үг хэллэгт зориулсан үйлдэлтийн утгазүй (procedural semantics)-тэй. Хэрэглэгч роботоос блокийн хэсгийг удирдахыг хүсэх, хэлбэржүүлэхийг хүсэх, шалтгааныг нь асуух зэрэг үйлдэл хийх боломжтой. Жишээлбэл:

- Pick up a red block.
- OK.
- Grasp the pyramid.
- I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
- Find a block which is taller than the one you are holding and put it into the box.
- BY 'IT', I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING. Г.М.

1980-1990 –ЭЭД ОН

Энэ үеийг **Эмпиризмийн эгэл** (Return of Empiricism) гэж болно. Ярианы тэмдэглэгээ хадах (speech tagging), хэлзүйн задлал хийх (parsing), үгийн утга онтусгах (word sense disambiguation)-ад зориулсан магадлалын арга барилын хэрэглээ гарч ирж, хэлийг ойлгох, сурах зорилтод зориулсан загвар, тэмдгийн, магадлалын загварын харьцуулал бий болов. Технологийн хөгжлийн дүнд дискний зайн бэрхшээл шийдэгдэж, машин уншиж болох бичвэр боловсруулагдаж, АНУ-ын хөрөнгө оруулалт өгөгдлийн хэмжээг онцолж, томоохон хөмрөгүүд байгуулсан. Үүнд, Америкийн Үндэсний Хөмрөг³², Британийн Үндэсний Хөмрөг³³, Ворднет³⁴ Вэб³⁵-ийг дурдаж болно.

1990-ээд онд тоо бүртгэлийн БХБ (Statistical NLP) бий болоход яриа таних нь хангайттай хэмжээний өгөгдөл, энгийн статистик аргууд хэрэгтэйг харуулсан, АНУ-ын хөрөнгө оруулалт ярианд суурилсан хэрэглэгчтэй харилцах хэсэг болон мэдээлэл гаргаж авахыг онцолсон, их хэмжээний тоон хөмрөг бий болсон зэрэг хүчин зүйл нөлөөлжээ. Тоо бүртгэлийн БХБ нь хэлзүйн дүрмээр үүсгэхэд эргэлзээтэй урт өгүүлбэр гарч ирэх, хэдэн мянгаас хэдэн сая хүртэлх задлал хийх зэрэг өмнө тулгарч байсан бэрхшээлийг шийдэхэд магадлалын болон тоо бүртгэлийн аргыг хэрэглэж эхэлсэн.

1990-Одоо

Яриа таних ба үүсгэх, зөв бичих зүйн болон хэлзүйн алдаа засах, машин орчуулга, мэдээлэл сэргээх зэрэг БХБ-ын хэрэгцээг хангах програм хангамжийн болон техник хангамжийн дэвшлүүд бий болсон. Хэрэглээний програмд стохастик болон тэмдгийн аргыг хослуулах зэрэг давуу тал гарснаар ЦХШ улам өргөн хүрээтэйгээр хөгжих болов.

³¹ SHRDLU гэдэг нэрийг англи хэлний хамгийн түгээмэл хэрэглэгддэг үсгийг давтамжаар нь дараалуулсан ETAOIN SHRDLU гэдгээс үүсгэжээ.

³² Америкийн Үндэсний Хөмрөг (The American National Corpus/ANC) 1990 оноос боловсруулж, бичгийн болон ярианы 22 сая үгтэй болсон хөмрөг. <http://americannationalcorpus.org/>

³³ Британийн Үндэсний Хөмрөг (The British National Corpus/BNC) Маш их сурвалжаас гаргаж авсан бичгийн ба ярианы хэлний жишээний 100 сая үгийн цуглуулга. 1991-1994 онд байгуулсан. <http://www.natcorp.ox.ac.uk/>

³⁴ Ворднет (WordNet) Англи хэлний үгийн сангийн өгөгдлийн сан (lexical database) юм. Миллер (George Miller, 1920-) -ийн удирдлагаар 1985 оноос хөгжүүлж эхэлсэн ба хэлшинжлэлч Феллбаум (Christiane D. Fellbaum) чухал үүрэг гүйцэтгэсэн. <http://wordnet.princeton.edu/>

³⁵ Дэлхий даяарх вэб (world wide web/www) Тим Бернерс Ли (Tim Berners Lee, 1955-) 1989 онд бүтээсэн. 2000 оноос хөмрөг болгон ашигласан нь ЦХШ-ийн хөгжилд чухал түлхэц болж буй.

Ийнхүү хоёр хэлний хооронд орчуулга хийх хүчин чармайлтаар эхэлсэн ажил цахим тооцоолуурын тусламжтай хэлийг хэрхэн илэрхийлж, хэрхэн боловсруулах, хөмрөгийг ухаалгаар боловсруулах, хэлийг үүсгэх, задлах алгоритмуудыг хөгжүүлэх судалгааны бүхэл бүтэн талбар болж, цахим хэлшинжлэлчид төрж гарах болжээ.

3. Цахим хэлшинжлэлийн салбарууд

- ЦХШ-ийг судлагдахуун нь яриа эсвэл бичвэр болж байгаагаас хамааран
- Ярианы хэлийг ойлгох, ярианы хэлнээс бичвэр үүсгэхтэй холбоотой яриа таних (speech recognition), яриа бий болгох (speech synthesis) салбарууд.
 - Хэлийг тусгаар хэсгүүдэд задлах буюу нэгтгэхтэй холбоотой хэлзүйн задлал хийх (parsing) ба хэлзүйн дагуу бүтээх (generation) салбар гэж хуваан үзэж болох юм.

ЦХШ-ийн дэд салбарууд:

- Байгалийн хэлний боловсруулалт (Natural language processing)
- Тооцооллын утгазүй (Computational semantics)
- Хөмрөгийн хэлшинжлэл (Corpus linguistics)
- Хэлзүйн задлуурын зохиомж (Design of Parsers)
- Тэмдэглүүрийн зохиомж (Design of Taggers)
- Машин орчуулга (Machine Translation)

Эдгээр салбарын судалгааны үр дүн болох хэрэглээний програмууд (applications)-ын жишээ [Claire Gardent 2003, Frank Richter 2005/2006, Shuly Wintner 2004]:

- Яриа таних (Speech recognition)
- Яриа бий болгох (Speech synthesis)
- Асуултад хариулах (Question answering)
- Харилцан ярих систем (Spoken Language Dialog Systems)
- Бичвэр хураангуйлах (Text Summarisation)
- Зөв бичлэг шалгах (Spelling checking)
- Хэлзүй шалгах (Grammar checking)
- Найруулга шалгах (Style checking)
- Мэдээлэл эрэх ба гаргаж авах (Information retrieval and extraction)
- Эх хэлний интерфейс (Natural Language Interfaces)
- Нэг хэлнээс нөгөө рүү орчуулах (Machine Translation)

Яриа таних (Speech recognition) нь яриаг бичвэр болгон хувиргахад чиглэнэ. Анхны яриа танигч 1952 оноос бий болсон [Davies 1952, pp.637-642]. Яриа таних систем (ЯТС/Speech recognition system)-ийн үзүүлэлт нь ихэвчлэн таних чадвар, хурдны тодорхой нөхцөлд биелдэг. Тэрхүү нөхцөл нь: ярианы онцлог бэлтгэсэн өгөгдөлтэй тохирч байх, ярьж буй хүн зохистой тааруулж ярих, шуугиангүй орчинд ажиллах юм.

Энэ төрлийн програм гэвэл, автомат орчуулга, автомат яриа танигч, бодит хугацааны дуу бичигч, хэрэглэгчийн интерфейст дуун команд таних, гар утас, дуудлага (хэл сурах програмд бий), видео тоглоом, ярианаас бичвэрт хөрвүүлэгч, агаарын замын хяналтын яриа танигч зэрэг болно. ЯТС ойролцоогоор 60-70% танина. Гол бэрхшээл нь хэлэхүйн явцаас үгийг тасалж авах чадвар, үгийн сангийн хэмжээ, хүний ярьж буй байдлаас хараат байдал, алдааг таньж илрүүлэх явдал.

Яриа нэгтгэх (Speech synthesis) Бичвэрээс ярианд хөрвүүлэгч (Text to Speech/TTS) нь бичвэрийг ярианд хувиргадаг, бусад систем авианы хөрвүүлэг (phonetic transcription) гэх мэт тэмдгийн хэлшинжлэлийн илэрхийллийг ярианд орчуулдаг. Нэгтгэсэн яриа нь өгөгдлийн сан хадгалагдсан бичигдсэн ярианы хэлхээсээр бүтээгдэж болно. Системүүд

хадгалсан ярианы нэгжийн хэмжээгээрээ ялгардаг. Яриа нэгтгэгчийн 2 чухал чанар нь хүний дуу хоолойтой адил байх (naturalness), ойлгомжтой байх (intelligibility) юм.

Маш эрт үед энэ тухай төсөөлөл байсныг дурдвал Герберт, Магнус, Бейконоос авахуулаад “ярьдаг толгой”-н домог байв. Краценштайн (Christian Kratzenstein) хүний дууны замын загварыг байгуулж (1779), Кемпелен (Wolfgang von Kempelen) дуу авианы механик ярих машин дүрсэлж (1791), Вийтстоуне (Charles Wheatstone) “ярьдаг машин” гаргаж (1837), Фабер (M. Faber) “Euphonia”-г байгуулав (1857). 1930-аад онд Белл лаборатори цахим яриа задлагч ба нэгтгэгч “VOCODER”-ийг хөгжүүлж, Дүдлей (Homer Dudley) “VODER” болгон шинэчилж, 1939 оны Дэлхийн Үзэсгэлэн худалдааны үзмэрт тавьсан. 1961 онд Келли (John Larry Kelly), Герстман (Louis Gestman) нар яриа бий болгоход “IBM 704”-ийг ашигласан нь Белл лабораторийн цуутай үйл явдал байв.

Асуултад хариулах (Question answering/QA) бол асуултад автоматаар хариулах явдал юм. Хариуг олохын тулд урьдаас бэлдсэн өгөгдлийн сан буюу бичиг баримтын цуглуулга хэрэглэх ба хөмрөгийн хайлтаас шууд хамааралтай. “АХ”-д түлхүүр үгэнд тулгуурлаж болно. Ийм “гүехэн” (shallow) арга хангалттай бус үед нэрлэсэн нэгжийг таних, холбоог илрүүлэх, өгүүлбэрзүйн өөрчлөл, үгийн утга онтусгах, логик хувиргал зэрэг хам сэдвийн боловсруулалтын “түн” (deep) аргыг ашиглана.

1960-аад оны АХС бол “Baseball”, “Lunar” юм. Зарим систем асуултад хариулах чадамжтай байснаас алдартай нь “SHRDLU”, “ELIZA”. 1970-1980-аад оны ЦХШ-ийн онолын хөгжил АХС-д зохистой нөлөөлсөн. Жишээ нь “Unix Consultant”, “LILOG”. 1990-ээд оны сүүлээс “Text Retrieval Conference” (TREC) тэмцээн зохиогдох болж, 2004 оны шилдэг систем 77%-ийн амжилт үзүүлжээ. 2007 онд “TREC” блогийн өгөгдлийг хөмрөг болгон оруулав. Блог бол бодит амьдрал дээрх өгөгдөл, хүмүүс зөв бичихэд анхаардаггүй тул алдаатай, замбараагүй байдаг. Одоо АХ-ыг вэб хайлт (web search)-ад нэгтгэх сонирхол өсөж, Гүүгле, Майкрософт хийж эхэлсэн. Сүүл үед “True knowledge”, “Wolfram Alpha” гэх мэт хариулт тооцоолдог хэрэгсэл гарч байна.

Харилцан ярианы систем (Spoken Language Dialog System) нь тодорхой бүтцийн дагуу хүнтэй ярилцах зорилготой, харилцахдаа бичвэр, яриа, дүрс, хүртэхүй, дохио болон бусад загварыг ашигладаг цахим тооцоолуурын систем юм. Гол ойлголт нь яриа таних, нэгтгэх, харилцан ярианы зохицуулалт (dialogue management) дээр тулгуурлах ба зохицуулагч нь ярианы байдал, бодлогыг удирдана. Хэрэглээний програм нь эшлэлийн сан, цагийн хуваарь гэх мэт “мэдээллийн” (Information services), банк, мөрий, нислэгийн захиалга гэх мэт “шилжүүлгийн” (Transaction services) үйлчилгээнд хэрэглэгдэнэ.

Ийм системд бэлтгэл үе шат байх боломжгүй, алдаа илрүүлэх нь хүндрэлтэй, хэрэглэгчийн санаачлага хязгаартай, хязгаарлагдмал нөхцөлд зохицуулалт хийдэг, байгалийн хэлийг заримдаа тааруухан ойлгох зэрэг байдал ажиглагдаж буй.

Бичвэр хураангуйлах (Text summarization/Automatic summarization) нь бичвэрийн товч хувилбарыг гаргах зорилготой. Үйл явцын дүнд жинхэнэ эхийн чухал хэсгүүд агуулагдаж үлдэх ёстой. Мэдээлэл хэт ачаалагдах нь дэс дараатай, зөв хөгжүүлсэн хураангуйлуур амин чухал гэдгийг илтгэдэг. Энэ нь төрөл бүрийн бичвэрийн уялдаа сайтай хураангуйг хийж чадах ба ингэхдээ хэмжээ, өгүүлбэрзүй, бичлэгийн хэв маяг зэргийг тооцох хэрэгтэй. Хураангуйлахад *хэсэглэх* (extraction), *товчлох* (abstraction) гэсэн хоёр хандлага бий. *Хэсэглэх* нь гол өгүүлбэр, догол мөр зэрэг хамгийн чухал хэсгийг хувилж тавьдаг бол *товчлох* нь баримтын хэсгээс найруулан бичих үйлдлийг

агуулдаг. Ерөнхийдөө товчлох нь хэсэглэхээсээ илүү үр дүнтэй боловч хүний хэл үүсгэх технологийн хэрэглээг шаарддаг тул хөгжүүлэхэд амаргүй.

Бүрэн автомат (Fully Automated), машины туслалцаатай (Machine Aided), хүний туслалцаатай (Human Aided) зэрэг төрлүүд бий. [Mani, 2001] Хураангуйлах технологийн хэрэглээний нэг жишээ бол "Google" гэх мэт хайлтын хөдөлгүүрүүд юм. Зөв бичлэг шалгуур (Spell checker) бол баримтын доторх зөв бичигдээгүй байж магадгүй үгийг тэмдэглэдэг програм. Энгийн ЗБШ нь толь бичигтэйгээ үг бүрийг тулгаж ажиллах ба үг олдоогүй бол алдаа гэж үзээд, байж болох хувилбарыг санал болгоно. Байхгүй үгийг нэмэх боломжтой, ингэснээр дараа удаа алдаа гэж үзэхгүй юм.

Анхны ЗБШ-ыг 1970-аад оны сүүлээр Жоржтауны хэлшинжээчдийн баг "IBM"-д зориулж хөгжүүлсэн ба "PC"-д зориулсан ЗБШ 1980 онд гарчээ. Өнөөдрийг хүртэл ЗБШ боловсронгуй болсоор, зарим нь энгийн хэлзүйн алдааг ч таних боломжтой болж байна. Зарим үед алийг нь буруу гэх вэ гэдэг асуудал гардаг. Жишээлбэл, хөмрөгийн хэлшинжлэлд тулгуурлавал, "baht" гэдгийг "bath" юм уу "bat" -ийн буруу бичлэг байх боломжтой гэж үзнэ. Гэтэл хэрэв бага хэрэглэгддэг ч Тайландын мөнгөний нэгж (baht) байвал яах вэ гэх мэт. Тиймээс сүүлийн үеийн судалгаа нь үгийн ойр орчныг мэдрэхэд онцгой анхаарч байна. Энэ нь зарим талаар толь бичиг баяжуулах ажлыг сулруулах боловч, олон үгийг таних боломж олгодог. Үүнд холбогдох хамгийн тохиромжтой, хамгийн алдартай жишээ бол Жерролд Зарын 1992 онд бичсэн шүлэг юм.³⁶

"Candidate for a pullet surprise"

"I have a spelling chequer.

It came with my pea sea.

It plane lee marks four my revue

Miss steaks aye can knot sea.

Eye ran this poem threw it

Your sure reel glad two no

Its vary polished inn it's weigh.

My checker tolled me sew." гэсэн эхний 8 мөрийг орчуулбал,

"Надад зөв бичлэгийн дөрвөлжин хээ бий.

Энэ миний вандуй тэнгистэй хамт ирсэн.

Энэ онгоц нөмөр газар миний дөрвөн үзмэрийг тэмдэглэдэг.

Залуу хатагтай хэрчим махнууд дандаа тэнгисийг зангидаж чаддаг.

Нүд энэ шүлгийг ажиллуулж түүнийг шидсэн.

Чиний итгэлтэй дамар баяртай хоёр үгүй

Үүний хувилбар дэн буудлыг жингээрээ өнгөлнө.

Миний шалгуур надаас оёж татвар авсан" гэсэн шиг болно.

Үг үгээр нь шалгадаг боловсронгуй бус зөв бичлэг шалгуур эндээс буруу бичсэн ямар ч үг олохгүй. Учир нь англи хэлэнд үнэхээр л *pea, sea, eye, steak, two, toll, sew* гэхчлэн үгс бий. Энэ ойлгохын аргагүй болгоод байгаа үгсийг дуудлага адилхан боловч хам сэдэвт тохирох утгатай, зөв бичлэгтэй үгээр нь сольж, орчуулгыг харгалзуулбал:

³⁶ Жерролд Х. Зар (Jerrold H. Zar) бол Хойд Иллинойс Их сургууль (Northern Illinois University)-ийн биологийн ухааны профессор байсан. Жинхэнэ бүрэн эхийг дараах сайтаас үзнэ үү.
<http://www.bios.niu.edu/zar/poem.pdf>

I have a spelling checker	(Надад зөв бичлэг шалгуур бий.
It came with my PC (Personal Computer)	Энэ миний ПиСи (PC)-г дагалдаж ирсэн.
It plainly marks for my review mistakes I can not see.	Энэ миний хараагүй алдаануудыг тэмдэглэж өгдөг.
I run this poem through it	Би энэ шүлгийг тэр шалгуураар оруулсан.
You're sure real glad to know.	Та олж мэдээд үнэхээр баярлах болно.
It's very polished in its way.	Энэ өөрийнхөө хэмжээнд их боловсронгуй.
My checker told me so.	Миний шалгуур надад тэгж л хэлсэн.)

Харин хам сэдэвтэй нь шинждэг шалгуур бол алдааг мэдрэх боломжтой. Одоогоор Голдин, Рот нарын алгоритм [Golding, Roth, 1999] хамгийн үр дүнтэй буюу 96% таних боломжтой. "MS Office 2007"-д хам сэдвийн мэдрэмжтэй ЗБШ гарч ирсэн.

Хэлзүй шалгуур (Grammar checker) бол бичвэр хэлний зүйн хувьд зөв байгаа эсэхийг шалгаж нотлох зорилготой бие даасан програм буюу програмын хэсэг юм. Эртний ХЗШ жинхэнэ хэлзүйн алдааг олохоосоо илүү цэг таслал, хэлбэрийг шалгадаг, програмын амин хэсэг нь мэргэжилтнүүдийн шалгаж, тааруу гэж үзсэн хэдэн зуу, хэдэн мянган үг хэллэгийн жагсаалт байв. Тэрхүү сэжигтэй гэж үзсэн хэллэгийн жагсаалт нь байж болох өөр хувилбарыг агуулна. Програм бичвэрийг өгүүлбэрээр нь задалж, толь бичигтэйгээ тулгаад, сэжигтэйг нь тэмдэглэж, өөр хувилбарыг харуулна.

1970-аад онд Юниксд багтаж байсан "Writer's Workbench" бол анхны систем. Өөр ийм програм гэвэл, 1981 онд "PC"-д зориулан гаргасан "Grammatik", "Punctuation and Style", "Correct Grammar", "RightWriter"-ийг дурдаж болно.

Жинхэнэ ХЗШ илүү цогц шинжтэй байх ёстой билээ. Хамгийн чухал хэсгийн нэг нь хэлний бүх үгийг, үг тус бүрийнх нь үгийн аймагтай хамт агуулсан баялаг үгийн сангийн хөмрөгтэй толь бичиг юм. ХЗШ бичвэр доторх өгүүлбэр бүрийг хайж олоод, улмаар толь бичиг дэх үг бүрийг олно. Дараа нь өгүүлбэрийг хэлзүйн дагуу задлахыг хичээнэ. Олон дүрэмд тулгуурлан програм үйл үгийн цагийн зохицол, тооны зохицол, үгийн дараалал гэх мэтчилэн алдааг илрүүлнэ. Мөн идэвхгүй хэвийн амаргүй хэрэглээ зэрэг найруулгын зарим алдааг ч илрүүлэх боломжтой. Өгүүлбэрийг задалсны дараагаар, зохистой хэлбэрээр нь дахин найруулж бичнэ гэсэн үг юм.

Мэдээлэл эрэх (Information Retrieval/IR) бол мэдээлэл, баримт, дээд түвшний өгөгдөл (metadata) хайх, холбогдох өгөгдлийн сан болон вэбийн хайлтыг судалдаг. Мэдээллийн хэт илүүдлийг багасгахад хэрэглэгдэнэ. Цахим тооцоолуураар мэдээллийн нэг хэсгийг хайх санаа Бушийн өгүүлэлд гарчээ³⁷. Анхны систем 1950, 60-аад онд, гарсан. Эхэндээ "Cranford collection" зэрэг бага хөмрөг дээр ажиллаж байв. 1970-аад онд өргөн хүрээтэй "Lockheed Dialog system" зэрэг ситем хэрэглээнд орсон. 1990-ээд оны сүүлээр хайлтын хөдөлгүүрийн олон онцлог шинж чанарын гүйцэтгэл нь ийм системээс гарах болов. Хайлтын хөдөлгүүрүүд бол мэдээлэл эрэх загвар, судалгаа, түүний хэрэгжилтийн хамгийн нийтлэг, магадгүй хамгийн тодорхой жишээ нь юм.

Мэдээлэл эрэх ажиллагаа нь хэрэглэгч хүсэлтээ (query) системд оруулснаар эхэлнэ. Баримтууд мэдээлэл эрэх системд шууд бусаар, дээд төвшний өгөгдөл байдлаар хадгалагддаг. Ихэнх мэдээлэл эрэх системүүд өгөгдлийн сан дахь хэрэглэгчийн

³⁷ **Буш** (Vannevar Bush, 1890-1974) аналог тооцооллын ажлаараа алдартай. Түүний "As We May Think" хэмээх өгүүлэл "Atlantic Monthly" сэтгүүлд 1945 онд хэвлэгдсэн.

хүсэлтэд объект хэр тохирч буйг илтгэсэн тоон үзүүлэлтийг тооцоолж үзээд, үүгээр нь объектыг эрэмбэлж, эхний эрэмбийн объектуудыг харуулна. [Singhai, 2001].

Мэдээлэл гаргаж авах (Information Extraction/IE) бол мэдээлэл эрэх (IR)-ийн нэг хэлбэр. Гол амжилт нь интернет зэрэг тодорхой бүтэцгүй буюу дээд төвшний өгөгдөлгүй хэлбэр, тодорхой бус бүтэцтэй машин уншиж болох баримт, тусгай сурвалжаас хам сэдэв болон утгын хувьд сайтар тодорхойлогдсон, ангилсан, бүтэц бүхий мэдээллийг автоматаар гаргаж авдаг явдал болно. Энгийн зорилгууд нь агуулгын булингарыг арилгах (remove content noise): зар сурталчилгаа гэх мэт хэрэггүй хэсгийг арилгах, нэрлэсэн нэгжийг таних (recognize named entity): хүний нэр, газрын нэр, хугацааны нэгж зэрэг нэрийг таних, нэр томьёо гаргаж авах (extract terms): өгөгдсөн хөмрөгт холбогдох нэр томьёог олох г.м

Сүүлийн үеийн мэдээлэл гаргаж авах технологийг зохицох чадвартай, уян хатан болгох хичээл зүтгэл нь сайтар зохион байгуулснаас аваад холимог, бараг чөлөөт гэж болох бичвэр хүртэлх олон төрлөөс мэдээлэл гаргадаг болж байна. Нээлттэй эх бүхий ийм програм, үйлчилгээнд “GATE”, “Open Calais”, “Mallet” -ийг дурдаж болно.

Машин орчуулга/МО (Machine Translation/MT) бол бичгийн болон ярианы ямар нэг бичвэрийг нэг хэлнээс нөгөө рүү орчуулах зорилготой салбар юм. [Hutchins, Somers, 1992] Түүхэн үүднээс ЦХШ-ийн эх үүдэл болсон салбар, мөн оруулж байгаа өгүүлбэр бүрийг задлан шинжлэх болон гарч буй өгүүлбэрийг үүсгэхтэй холбоотой бүхий л алхмыг хэлшинжлэлийн үүднээс шаарддаг онцлог нь хамгийн чухал хэрэглээний талбар болгодог. Тиймээс машинаар орчуулах нь ЦХШ-ийн мэдээллийг эрхшээх ба түгээх хүч тэнхээ шавхсан, цогц байх ёстой зорилтын нэг юм.

1954 онд “The Georgetown experiment” 60 орчим өгүүлбэр орчуулсан нь томоохон амжилт байсан. Мөн 1954 онд Эндрю Бүүт³⁸-ийн зохиосон “APEXC” машин англиас франц руу анхан шатны энгийн орчуулга хийж байв. Өмнө дурдсанчлан 1966 оны “ALPAC”-ын тайлангаар МО зогсонги төлөвтэй болсон юм. 1980-аад оны сүүлээс тооцооллын хүч нэмэгдэж, үнэ харьцангуй буурч эхэлснээр машин орчуулгын тоо бүртгэлийн загваруудыг илүүтэй сонирхох болж эхэлсэн.

Машин орчуулгын үйл явц:

1. Сурвалж хэл (source language)-ээрх бичвэрийн утгыг дүрэмд суурилсан хэлшинжлэлийн аргаар задалж кодчилон тайлна.
2. Завсрын төвшний тэмдгийн илэрхийллийг бүтээнэ.
3. Зорьсон хэл (target language) дэх тэрхүү утгыг дахин кодчилж хамгийн тохиромжтой хэлбэрээр орчуулга үүсгэнэ.

Энэ энгийн санагдах үйл явцын ард үнэндээ маш их адармаа байдаг. Эхлээд хэлийг ойлгохтой тулгарна. Сурвалж бичвэрийн утгыг бүрэлдэхүүнээр нь задлахын тулд, бичвэрийн бүх онцлогийг шинжлэх ёстой бөгөөд үүнд хэлзүй, утгазүй, өгүүлбэр зүй, хэвшмэл хэллэг гэх мэт сурвалж хэл болоод түүгээр хэлэлцэгчийн соёлын талаарх гүнзгий мэдлэг шаардагдана. Эргээд кодчилж орчуулахад зорьсон хэлнийх нь талаар сурвалж хэлнийхтэй адил мэдлэг хэрэгтэй. МО-ын дараах загварууд байдаг.

а. **Дүрэмд суурилсан МО** (Rule based MT)-д шилжүүлэгт суурилсан (transfer-based), хэл хоорондын (interlingual), толь бичигт суурилсан (dictionary-based) зэрэг загварууд багтана. Энэ хандлагад толь бичгийг голчилно.

³⁸ Бүүт (Andrew Donald Booth, 1918-2009) Соронзон бөмбөр санах ой (magnetic drum memory)-г зохиох ажлыг удирдаж байсан ба Бүүтийн үржүүлэх алгоритмыг зохиосноороо алдаршсан.

б. **Тоо бүртгэлд суурилсан МО** (Statistical MT): Хоёр хэлний хөмрөгт тулгуурлаж тоо бүртгэлийн аргаар орчуулдаг. Хөмрөгийн жишээнд Канадын парламентын илтгэл “Hansard”, европын парламентын илтгэл “Europarl”-ийг дурдаж болно. Ийм хөмрөгүүд байвал ижил төрлийн бичвэрийн орчуулга үр дүнтэй болно. Гэвч ийм хөмрөг маш ховор. Анхны тоо бүртгэлийн МО-ын програм бол “IBM”-ийн “CANDIDE”. “Google” “SYSTRAN”-ыг хэрэглэж байгаад 2007 оноос тоо бүртгэлийн орчуулга руу шилжсэн ба НҮБ-ийн 200 биллион үгтэй хөмрөг ашиглан орчуулгынхаа чадамжийг сайжруулж буй.

в. **Жишээнд суурилсан МО** (Example based MT/EBMT): Мэдлэгийн суурь болсон 2 хэлний хөмрөгийг хэрэглэдгээрээ онцлог ба мөн чанартаа аналог орчуулга тул тохиолдолд суурилсан бодлогын чиглэл³⁹-ийн хэрэгжилт гэдэг.

г. **Хосолмол МО** (Hybrid MT/HMT): Тоо бүртгэлийн болон дүрэм суурилсан аргыг хослуулдаг. “Asia Online”, “Systran” зэрэг компаниуд үүнийг гаргаж ирж байна.

Гол асуудал

МО-ын хамгийн чухал асуудал буюу бэрхшээл бол **үгийн утгыг онтусгах** (word sense disambiguation) юм. Энэ нь, үг хэд хэдэн утгатай байж болохоос өгөгдсөн хам сэдэвт чухам аль утга, хэлбэр тохирч байгааг онож тодорхойлох зорилготой хэлний бүх түвшинд хамаарах судалгаа болно. Үүнийг анх 1950-аад онд Бар-Хиллел⁴⁰ хөндөн гаргаж, нэвтэрхий тольгүйгээр машин нэг үгийн олон утгын ялгааг ялгах боломжгүйг онцолжээ [Bar-Hillel 1960]. Эхэн үеийн аргууд нь өгөгдсөн хам сэдэв дэх утгыг сонгож тодорхойлж байсан. Жишээ нь, “*The astronomer saw the star*”, “*The astronomer married the star*” гэсэн өгүүлбэрийн “star” нь тэнгэрийн од уу, нэрт жүжигчин үү гэдгийг танихдаа хүний оролцоог хэрэглэж байсан.

Өнөөдөр үүнд чиглэсэн олон хандлага гарч буйг ерөнхийд нь гүн ба өнгөц гэсэн хоёр хэсэгт хуваан үзэж болно. Өнгөц хандлага (shallow approach) нь олон утгатай үгсийн ойр орчны үгсийг тооцох тоо бүртгэлийн аргыг хэрэглэж, гүн хандлага (deep approach) нь үгийн тухай дэлгэрэнгүй мэдлэгийг авч үздэг. НҮБ-ийн орчуулагч асан Пирон машин орчуулга мэргэжлийн орчуулагчийн ажлын 25%-ийг автоматчилж, хэцүүхэн 75%-ийг нь үлдээх ба энэ хамгийн их цаг авдаг хэсэг нь сурвалж бичвэр дэх олон утгыг зорьсон хэлэнд оноох явдал болохыг өгүүлсэн байдаг [Claude Piron, 1994]:

“Орчуулагч яагаад 5 хуудас орчуулахын тулд 1, 2 цаг биш бүтэн ажлын өдөр зарцуулдаг вэ? ...Дундаж бичвэрийн 90% орчим нь энгийн байдаг. Харамсалтай нь, дахиад 10% үлддэг. Энэ хэсэг л ажлын 6, түүнээс их цагийг шаарддаг юм.”

Гүн хандлага нь утга онооход хэрэгтэй бүх судалгааг шаардана. Гэвч хиймэл оюун ухааны ийм дээд төвшинд хүрэх хараахан болоогүй байна. Тиймээс одоогоор өнгөц хандлага илүү үр дүнтэй юм. Өнгөц хандлага утгыг таамаглаж тавих ба үнэндээ ихэвчлэн буруу байдаг тул “тохирох утгыг хэрэглэгчээс асуух” үйлдлийг багтаадаг.

4. ЦХШ-д ХШ холбогдох нь

Хэлийг цахимын орчинд боловсруулах, түүнд суурилсан хэрэглээний програм бүтээхэд тухайн чиглэлээсээ хамаарч хэлний бүхий л төвшний мэдлэг оролцдог.

Авиазүй нь голчлон хэлний авиаг таних, улмаар тэдгээрийг зөв дараалалд оруулахад хэрэгтэй. Үгийг авиа болгон буулгах буюу яриа нэгтгэх (speech synthesis),

³⁹ **Тохиолдолд суурилсан бодлого** (case-based reasoning) бол шинээр тулгарч буй асуудлыг өмнө тохиолдсон асуудлын шийдэлд тулгуурлан шийдэх үйл явц юм.

⁴⁰ **Бар-Хиллел** (Yehoshua Bar-Hillel 1915-1975) МО, хийсвэр хэлшинжлэл, мэдээлэл эрх талбарт санаачилсан ажлуудаараа алдаршсан. Хожим нь ерөнхий, бүрэн автомат, дээд төвшний МО байна гэхэд эргэлзэж буйгаа илэрхийлсэн [Bar-Hillel, 1960, p.158-163].

авиаг үг болгон буулгах буюу яриа таних (speech recognition) салбарын судалгаанд үндсэн суурь болж өгдөг. Энэ салбарын бэрхшээл нь үгийн хязгаарыг таних явдал юм.

Үгзүй нь хам сэдэв дэх үгийн харьцаа, хэлбэрийн талаарх мэдээллийг гаргаж өгнө. Энэхүү мэдлэгт тулгуурлан үгийг язгуур, залгавар гэхчлэн бүтцээр нь задлахын зэрэгцээ эсрэг үйлдэл болох үгийн зохистой хэлбэрийг үүсгэх үйлдлийг гүйцэтгэдэг.

Өгүүлбэрзүй нь үг, хэллэгийн хоорондын бүтцийн харилцааг илрүүлж бүтэц бүрэлдэхүүнийг тодорхойлно.

Хэлзүй нь хэлний өгүүлбэрийг тодорхойлоход хэрэгтэй. Эдгээр нь өгүүлэгдэхүүн, өгүүлэхүүний зохицол гэх мэт дүрэм шалгуур, өгүүлбэрийн утгыг гаргах, хэлзүйн дагуу өгүүлбэр үүсгэх суурь болдог.

Утгазүй нь үгийн утга, түүний хам сэдэвтэй харилцах холбоог тодорхойлж өгдгөөрөө чухал юм. Энэ мэдлэгт тулгуурлан, үгийн сангийн ангилал нэгжийн ойролцоо, төрөл, эсрэг, ижил нэрсийг тооцох үгийн утгыг онтусгах, машинаар орчуулах үед орлуулга хийх гэхчлэн олон чухал үйлдлийг гүйцэтгэнэ.

БХБ-ын модул ба хэлшинжлэлийн мэдлэг

Яриа таниур (Speech recognizer) ↔ Авиазүй (Phonology)

Үгзүйн задлуур (Morphological analyser) ↔ Үгзүй (Morphology)

Үгийн сан тэмдэглүүр (Part of Speech Tagger) ↔ Өгүүлбэрзүй (Syntax)

Хэлзүйн задлуур (Parser) ↔ Өгүүлбэр зүй, Үгзүй, Утга зүй (Syntax, Morphology, Semantics)

Онтусгах (Disambiguation) ↔ Утгазүй, ярилцлага, хэрэглээзүй (Semantics, Discourse, Pragmatics)

Эцэст нь тэмдэглэхэд, хэлшинжлэлийн мэдлэгүүд тухайн системийн нэг буюу хэд хэдэн модулийн мэдлэгийн суурь болон кодчилогдох байдлаар системд шингэж, утгын илэрхийллийг хэрэглэгчийн оруулсан өгөгдлөөс бүтээж, хариу үйлдлийг гаргаж үзүүлэх зарчимтай. ЦХШ-д үгзүй, өгүүлбэрзүй, зөв бичих зүй, хэлзүйн дүрэм, утгазүйн мэдээлэл, хөмрөг гэх мэт иж бүрэн өгөгдөл шаарддаг. Хангалттай өгөгдөлтэй үед програм үр ашигтай ажиллана. Тиймээс харилцан уялдсан арга барил, өгөгдлийг бүтээх чухал. Жишээлбэл, тоо бүртгэлийн арга (statistical methods)-д зориулсан том хөмрөг хэлзүйд суурилсан үед онцгой үүрэгтэй бус бол хэлзүйд суурилсан арга (grammar-based methods)-д хэлшинжлэлчийн хэлзүй нь сайтар зохимжлох чадвар гол болох гэх мэт.

ЦХШ-ийн ололт нь хүний хэлийг хүн шиг ойлгон боловсруулж, бүтээх цахимын тогтолцоо бий болгосонд оршино. Дэлхийн улс орнууд хэл, бичгээ цахим орчинд боловсруулах, зүй тогтлыг нь таниулахаар хүч хөдөлмөр, хөрөнгө зарцуулдаг болсон тул энэ салбарын судалгаа улам боловсронгуй болсоор байна.

Ашигласан лавлах хэрэглэгдэхүүн

Шагдарсүрэн Ц., 2007. *Эх хэлээ эвдэхгүй юм сан*, ISBN 99929-53-15-2, Улаанбаатар, 215 тал.

Alain Colmerauer 1969. *Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur*, Mimeo, Montral.

Amit Singhai 2001. *Modern Information retrieval: A Brief Overview*, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24 (4), pp. 35-43.

Andrew R. Golding, Dan Roth 1999. *A Winnow-Based Approach to Context-Sensitive Spelling Correction*, Machine Learning, Vol. 34, Springer Netherlands, pp.107-130.

Barbara Grosz, Candice L. Sinder 1986. *Attention, Intentions, and the Structure of Discourse*, Computational Linguistics 12, pp.175-204.

Carl Pollard, Ivan Sag 1994. *Head-Driven Phrase Structure Grammar*, University of Chicago press.

Claire Gardent, 2003. *A Brief Introduction to Computational Linguistics*, Tutorial on Computational Linguistics, (Tbilissi), p.110.

- Claude Elwood Shannon, 1951. *Prediction and Entropy or Printed English*, Claude elwood Shannon: Collected papers, IEEE Press, 1993, ISBN 0-7803-0434-9, pp.69-85.
http://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf
- Claude Piron 1994, *Le d%ofi des langues Du gâchis au bon sens*, Paris, L'Harmattan, p.336.
- Darrel Ince, 2008. *Oxford Monsudar Dictionary of The Internet - Интернетийн тайлбар толь бичиг*, Oxford University Press, Monsudar Publishing, Ulaanbaatar, p.432.
- Davies , K.H., Biddulph, R. and Balashek, S. (1952) *Automatic Speech Recognition of Spoken Digits*, J. Acoust. Soc. Am. 24(6) pp.637 - 642
- Frank Richter, 2005/2006. *Introduction to Computational Linguistics*, NLP, Intro, <http://www.ba.umist.ac.uk/public/departments/registrars/academicoffice/uga/lang.htm>
- George Miller, 2001. *Ambiguous Words*, Released: March 22, 2001, iMP Magazine: http://www.cisp.org/imp/march_2001/miller/03_01miller.htm
- Gerald Gazdar, Ewan H. Klein, Geoffrey K. Pullum, Ivan A. Sag 1985. *Generalized Phrase Structure Grammar*, ISBN 0-674-34455-3, Harvard University Press.
- Inderjeet Mani 2001. *Automatic Summarization*, ISBN 1-58811-06-5
- John W. Hutchins, Harold L. Somers, 1992. *An Introduction to Machine Translation*, London, Academic Press, ISBN 0-12-362830-X <http://www.hutchinsweb.me.uk/IntroMT-TOC.htm>
- John W. Hutchins 1996. *ALPAC: the (in)famous report*, Originally published in: *MT News International* 14 (June 1996), 9-12. Reprinted in: *Readings in machine translation*, ed. S.Nirenburg, H.Somers and Y.Wilks (Cambridge, Mass.: The MIT Press, 2003), 131-135.
- John W. Hutchins (1997), 'From first conception to first demonstration: the nascent years of machine translation, 1947-1954. A chronology.' *Machine Translation* 12, 195-252
- John W. Hutchins 2005. *The first public demonstration of machine translation: the Georgetown-IBM system*, 7th January 1954
- Joseph Weizenbaum, 1966. *ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine*, Communications of the Association for Computing Machinery 9, pp.36-45.
- Joseph Weizenbaum, 1976. *Computer Power and Human Reason: From Judgment To Calculation*, San Francisco: W. H. Freeman, 1976; ISBN 0-7167-0463-3
- Martin Volk, 2004. *Introduction to Natural Language Processing*, Stockholm University, <http://www.umiacs.umd.edu/~bonnie/courses/cmssc723-03>
- Noam Chomsky, 1956. *Three models for the description of Language*, IRE Transactions on Information Theory (2), pp.113-124. <http://www.chomsky.info/articles/195609.pdf>
- Orhan Zeynep, Altan Zeynep, 2005. *Determining effective features for the word sense disambiguation in Turkish*, Journal of Electrical and Electronics Engineering, Vol. 5, Num.2, 2005 (1341-1352) Istanbul University.
- Peter Toma, 2000. "From Serna to Systran", in *Early Years in Machine Translation*, ed. J. Hutchins.
- Ralph Grishman, 1986. *Computational Linguistics: An Introduction*, Cambridge University Press.
- Richard Montague 1970a. *Universal grammar*, Theoria, 36 (1970), pp. 373-398.
- Richard Montague, 1970b. *English as a Formal Language*, Linguaggi nella societa e nella tecnica, Bruno Visentini (ed.), Mailand, pp.189-223.
- Richard Montague, 1973. *The Proper Treatment of Quantification in Ordinary English*, Approaches to NL, Jaakko Hintikka, Julius Moravcsik, Patrick Suppes (eds.), Dordrecht, pp. 221-242.
- Shuly Wintner, 2004. *Hebrew computational linguistics: Past and future*. Artificial Intelligence Review, 21(2):113-138.
- The Association for Computational Linguistics, <http://www.aclweb.org>
- William N Locke, and Booth, A. Donald, eds. (1955) *Machine translation of languages: fourteen essays*, Cambridge, Mass.: M.I.T.Press. <http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf>
- Yehoshua Bar-Hillel, 1960. *The Present Status of Automatic Translation of Languages*, Advances in Computers, Vol. 1, p.91-163. <http://www.mt-archive.info/Bar-Hillel-1960.pdf>

- York Wilks. 2005. *The History of Natural Language Processing and Machine Translation*, In Encyclopaedia of Language and Linguistics, Kluwer: Amsterdam.
- Zellig Harris, 1951. *Methods in Structural Linguistics*, University of Chicago Press, xvi, pp.384.
- Zellig Harris, 1988. *Language and Information*, Columbia University Press, New York.
- Zellig Harris, 1989. *The Form of Information in Science: Analysis of an immunology sublanguage*, Springer.
- Zellig Harris 1991. *A Theory of Language and Information: A Mathematical Approach*, Oxford University Press.

SUMMARY

Computational Linguistics/CL is an interdisciplinary field dealing with the statistical and rule-based modeling of natural language from a computational perspective. CL originated with efforts in the United States in the 1950s to use computers to automatically translate texts from foreign languages, particularly Russian scientific journals, into English. When machine translation failed to yield accurate translations right away, automated processing of human languages was recognizing far more complex than had originally been assumed. In order to translate one language into another, it was observed that one had to understand the grammar of both languages, including both morphology and syntax. In order to understand syntax, one had to also understand the semantics and the lexicon, and even understand something of the pragmatics of language use. Thus, what started as an effort to translate between languages evolved into an entire discipline devoted to understanding how to represent and process natural languages using computer.

It is a newly developing field in Mongolian Linguistics. So, the article surveys sixty years of work in computational language processing. The article has the following sections:

- What's the CL?
- A brief history of CL
- Subfields and Applications
- CL and Linguistics

The main goal of the field is

- to build psychologically adequate models of human language processing capabilities on the basis of knowledge about the way in which humans acquire, store, and process the language.
- to build functionally correct models of human language processing capabilities on the basis of knowledge about the world and about language elicited from people and stored in the system.