G2P Ангилах арга: Автоматаар ангилан ялгах арга буюу Монгол бичгийн дүрсийг ангилан кодлох арга

G2P Sorting: An Automated Natural Sorting Method for Graphetically Encoded Mongolian

Зохиогч: Шэн Илэй
Author : SHEN Yilei

Огноо: 2018 оны 03 сарын 24
Date: 2018-03-24

G2P ангилах аргачилал буюу автоматаар ангилан ялгах арга буюу Монгол бичгийн дүрсийг ангилан кодлох аргачилалыг энэхүү баримтанд тусган харуулсан болно.
G2P sorting, as an automated natural sorting method for graphetically encoded Mongolian is presented in this document.

1. Монгол бичгийг ангилан ялгах конвенц
Mongolian sorting convention

1.1 Ангилчлах аргыг задлан харуулах нь.
Unfolded sorting

Хэдийгээр Монголчууд уугуул иргэддээ Монгол хэл зааж сургахдаа ерөнхийдөө үелэх хэлбэрийн аргачилалаар зааж сургадаг боловч, орчин үеийн Монгол хэлний толь бичиг нь цагаан толгойн дарааллаар ангилж оруулсанбайдаг. Гол төлөв зохиосон цагаан толгойн дараалалаар ангилсан байдал нь: Although Mongolian is generally taught to native Mongolians in a syllabic or semisyllabic way, modern Mongolian dictionaries are sorted alphabetically. The commonly assumed alphabet (unfolded alphabet) used in sorting is

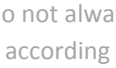
a ·e ·i ·o ·u ·ö ·ü ·ë ·n ·ŋ ·b ·p ·x ·g ·m ·l ·s ·š ·t ·d ·č ·ǰ ·y ·r ·w ·f ·k ·c ·z ·h ·ř ·ɬ ·ẑ ·ĉ
маш цөөн тооны үндсэн хувилбаруудтай. Үүнд:
with only a few minor variations:

- ë (ᠡᠶ) нь биеэ даасан үсэг эсвэл (ᠢ) хувилбарын нэг гэж авч үздэг.
  ë (ᠡᠶ) is treated either as an independent letter or as a variant of e (ᠢ);
- ŋ (ᠩ) нь мөн л биеэ даасан үсэг эсвэл (<ᠨ, ᠭ/ᠺ>) хувилбарын нэг болгон авдаг.
  ŋ ( ) is treated either as an independent letter or as a letter sequence ng (<ᠨ, ᠭ/ᠺ>).

Энэхүү Монгол цагаан толгойн үсгийн тэмдэглэгээ нь тод бус байдлаар тодорхойлогддог.Мэдээжээр, энэ тохиолдолд цагаан толгойн үсгийн хэд хэдэн бүлэг үсэг байдаг бөгөөд энэ нь тодорхой тохиолдлуудад ижилхэнбичсэн (a / e / n, ë / w, o / u / ö / ü, x / g, t / d гэх мэт) хувилбарууд нь өөр хоорондоо адил байдаг. Мөн үсгийн өвөрмөц шинж чанарууд нь ерөнхийдөө ярианы хэллэгийн дуудлаган дээр суурьлаж харьцуулсан байдаг. Гэсэн хэдий ч, хувь хүмүүс тухайн үсгийг тодорхойлж чадахгүй бэрхшээлтэй тулгардаг байна.
This Mongolian alphabet is characterized by its opaque identification of letters. As is known, there are several groups of letters in this alphabet which share the same written forms in certain cases (a/e/n, ë /w, o/u/ö /ü , x/g, t/d, etc.), and the distinctions of letters are established generally on the contrasts in colloquial pronunciations. However, individual cases of identifying letters can be problematic:

- Түүхийн явц дах авиа зүйн өөрчлөлттэй холбоотойгоор,зөвхөн ярианы хэл болон дуудлага дээр тулгууралсан үсгийг тодорхойлоход боломжгүй байдаг.Авиа зүй буюу фонологийн мэдлэгээс гадна, зөв бичих дүрэм, үгийн гарал зүй, бичигийн үсэг болон хэлний үг зүйн бүтэц ч мөн үүнд хамааралтай байдаг. Тиймээс хүн бүрийн гаргаж байгаа тодорхойлолтууд нь үргэлж өөр хоорондоо ялгаатай байх ба санал зөрөлддөг учраас хүн хүний гаргасан тодорхойлолтууд өөр өөр байдаг. Жишээ нь: "ᠠᠵᠢᠯ" гэдэг

Монгол үгийг бичихдээ čino_a, činu_a, эсвэл činw_a гэж бичигдэж байгаа хувилбар, шалгуурууд нь өөр хоорондоо зөрчилдөж хүн бүр өөр өөр тодорхойлж бичиж болдог.

Due to drastic phonological changes in history, identifying letters relying merely on colloquial pronunciation is not always possible. Besides phonological knowledge, one's orthographical, etymological, and morphological knowledge of the language/script also take part. Therefore it is quite often that one's identification may differ widely from another's as these criteria do not always converge. For example, the Mongolian word  "wolf" may be identified as čino_a, činu_a, or činw_a according to different criteria that contradict each other.

- Зөвхөн фонологийн буюу авиа зүйн мэдлэгийн хүрээнд, янз бүрийн нутгийн хэл, аялагаар ярьдаг хүмүүс өөрсдийнхөө орон нутгийн дуудлагуудын дагуу үсгээ тогтоох хандлагатай байдаг. Хэдийгээр стандарт дуудлагад бидний ярилцлагыг хязгаарласан ч гэсэн, Хятад болон Монгол улсын стандарт дуудлагууд нь олон талаараа өөр хоорондоо маш их ялгаатай болохыг тэмдэглэх нь зүйтэй. Гэсэн хэдий ч, өнөөгийн Хятад улс дахь Монголын стандартчилагдсан дуудлага нь Хятадад байгаа Монгол хүмүүсийн дунд өргөн хэрэглэгдэхгүйгээс гадна хамгийн оновчтой авиа зүй хэрэглэгддэгүй байна.

  As for phonological knowledge alone, people of various dialectal backgrounds tend to identify letters according to their local pronunciations. Even if confining our discussion to standard pronunciations, it should be noted that the standard pronunciations of China and Mongolia differ in many aspects. Worse still, the standard Mongolian pronunciation of China is an idealized phonology which is native to nowhere and is poorly popularized in China today.

Үүний үр дагавар нь, ангилчлалыг задлан харах үйл явц нь мөн эргэлзээтэй байна гэсэн үг.

As a consequence, the usability of unfolded collation is questionable.

- Нэг үг нь яг тодорхой үсгийн дараалалтай байх ба толь бичигт үсгийн дараалалын дагуу бичигдсэн болон зохиогч нь тухайн үгийг өөр өөр тэмдэгллэгээгээр тэмдэглэсэн тохиолдолд, тус үгийн үсгийн жинхэнэ дараалалыг олоход бүр ч их хэцүү байдаг.

  One word may be identified as various letter sequences and thus sorted variously among dictionaries, and one will not be able to find the entry when they has a different identification of the underlying letters from that of the dictionary compiler.

- Үгийг судлах явцад тухайн үгийн зөв дуудлагыг мэдэхгүй бол, толь бичигт оруулахын тулд зөв бичих дүрмийг нэг бүрчлэн эхнээс нь үзэж судлах шаардлагатай болно. Гэсэн хэдий ч, бичигдсэн нэг үгийг 4-өөс доошгүй хэлбэрээр унших нь элбэг тохиолддог. Жнь: talai, telei, dalai, and delei.

  When encountering a word they does not know how to pronounce correctly, one will have to go over every orthographically possible reading in order to find it in a dictionary. However, it is quite common for a single written form to have no less than four possible readings like talai, telei, dalai, and delei.

1.2 Ангилчлах аргыг нэгтгэх нь

Folded sorting

Гэсэн хэдий боловч, Монгол бичгийг ангилчлан нэгтгэхэд тулгарч байдаг асуудлууд ба дээр дурдсан орчин үе нь өмнөх үеийнхээс зайлсхийхэд хүрч байна. Хэний зарчим нь хамгийн энгийн зөв болох, нэг үеийн бүрэлдэхүүн, (эхэнд, дунд, адагт г.м.), нэг хэлбэр нь зөвхөн нэг үсгэн дотор багтаах, өөрөөр хэлбэл ангилах үйл явц нь үелэх бүтэц болон бичих аргачилал нь тодорхой болсон тохиолдолд бүрэн тодорхойлогдож байна гэсэн үг. Энэхүү аргыг "нэгтгэх арга" гэж нэрлэнэ. Учир нь ангилалын явцад үсэг болон үгүүд нь адил бичигддэг хэдийч өөр утгыг агуулгыг ялгаж ангилахад хэрэглэгддэг. Доорхи хэсэгт:

Nevertheless, there has been another approach to Mongolian sorting dating back to pre-modern times which can evade these above-mentioned shortcomings, whose principle is quite intuitive: for each component of the syllable (onset, nucleus, or coda), one shape is subsumed under only one letter, which means the sorting behavior for a specific written form is fully determined

as long as the syllable structure is known. This approach is termed here as “folded” because a syllabary/alphabet that folds homographic written forms is adopted in sorting. In this folded approach,

- (᠊᠊᠊) гэдэг үгийг telei and dalai гэж ангилж бичсэн байна. Учир нь энэ 2 адилхан бичигдсэн хэдий ч хоёрдмол утгыг хамгийн ихээр үгүйсгэсэн байна. telei and dalai (᠊᠊᠊) are sorted like *talai because they are written alike, which has maximally eliminated ambiguous reading;
- (᠊᠊᠊) гэдэг үгийг *bay_a, биш харин *bai_a (cf. bai (᠊᠊᠊)) ангилж бичсэн байна. Үечлэх бүтэц нь хэл шинжлэлийн хамгийн чухал ач холбогдолтой ангилалд багтдаг.
  bey_e (᠊᠊᠊) is sorted like *bay_a, not *bai_a (cf. bai (᠊᠊᠊)), where the syllable structure is retained to make the sorting linguistically significant.

Ангилчлах аргыг нэгтгэхэд үүсэх санал зөрчилдөөний улсаас хоёр өөр үзэл бодол гарах магдлалтай. Үүнд:
Arguments against folded sorting may come in two aspects:

- Энэ нь олон жилийн туршид хэрэглэгдэж ирсэн ялгах ангилалын нийтлэг практиктай зөрчилддөг.
  It goes against the common practice of unfolded sorting which has been established for several decades.
- Эдгээр ангилан нэгтгэх арга нь хэрэглэгчдэд цээжлэхэд хэцүү байдаг тул эдгээр нь ярианы хэл болон дуудлага дээр тулгуурлаагүй байдаг.
  The folded letters are difficult for users to memorize because they are not based on colloquial pronunciations.

Миний батлан харуулах гэж байгаа ангилчлах арга мөн хоёр талтыг тусган харуулж байна. Үүнд
My defense of folded sorting also comes in two aspects:

- Өнөөдөр өргөнөөр хэрэглэгддэг ангилан ялгах аргачилал нь уламжлалт аргыг сайн тогтоож өгөөгүй.
  The unfolded sorting widely adopted today itself is not a well-established tradition.
  - Үечлэн ангилах хэлбэрээс цагаан толгойн ангилах хэлбэрлүү шилжих үйл явц зөвхөн зуун жилээс дээшгүй байна.
    The shift from syllabic sorting to alphabetical sorting took place only no more than a century ago.
  - ŋ тэмдэгийг Хятад хэлээр хэвлэгдсэн толь бичиг дээрөөр ялгаатай байдлаар ангилан бичсэн байдаг.
    There are still discrepancies in unfolded sorting like how to arrange ŋ in dictionaries published in China alone.
  - Хятад хэлний стандарт доторх Монгол хэлний ялгах стандарт (GB / T 30851-2014) нь өргөн цар хүрээг хамарсан хэдий ч Хятад хэлний скриптүүдийн стандартуудтаас олон талаараа зээлж авсан байдаг.(GB / T 32912-2016 г.м). Илүү нарийн төдийгүй, ангилах стандартын текстийн тайлбар нь тэдгээрийн харьцуулах хүснэгттэй өөр хоорондоо мөн зөрчилддөг.
    The Chinese standard of Mongolian sorting (GB/T 30851-2014) contradicts widely adopted sorting practices and other Chinese standards of the script (e.g., GB/T 32912-2016) in many aspects. More ironically, the text description of the sorting standard contradicts with its collation table.
  - Хэрэглэгчид удаан хугацаагаар үсэгний ангилалаас үүсэн тогтворгүй байдал дээр дассан байна.
    Users have long been accustomed to the sorting instability resulting from fickle letter identification.

- Ангиллах аргыг нэгтгэх нь хэрэглэгчидэд хэрэглэж, эзэмшихэд илүү хялбар байдаг.
  Folded sorting is much easier for users to master.
  - Нэгтэгсэн үсэгнүүд нь цээжлэхгүй, харин шууд бичсэн хэлбэрээс авдаг.
    Folded letters are not memorized but derived directly from the written form.
  - Ангиллалыг нэгтгэх арга нь үгийн гарал зүй, үг зүйн бүтэцийн хувилбаруудын хувьд ангилчлан задлах аргатай адилхан байх шаардлаггүй. Зөвхөн ерөнхий зөв бичих дүрэм нь өгөгдсөн хэлбэрээс ангиллан нэгтгэх дарааллыг тодорхойлох бараг л хангалттай байдаг хэдий ч ярианы хэллэг хоёрдмол утгатай бол дуудлагаар ярих нь тустай байж болох юм.

Unlike unfolded sorting, etymological or morphological knowledge is no longer necessary in folded sorting. Basic orthography alone is almost enough to determine the folded letter sequence from a given shape, though colloquial pronunciation might be marginally helpful in case that syllabification ambiguity arises.

- Олон нийтэд зориулсан тоон текстэнд алдаа бичиж тэмдэглэх үед, хэрэглэгчид бичиж байхдаа нэгтэгсэн цагаан толгойн хэсгийг хэсэгчилсэн гэж үздэг.
- Typing errors in digital text contributed by the public shows that users have partially assumed a folded alphabet in typing.

Эцэст нь дүгнэж хэлэхэд, тодорхой бус талыг ангилан ялгахгүйн тулд эрэмбэлэх боломжгүй юм.
In the end, the conclusion is that in no definite aspect is folded sorting inferior to unfolded sorting.

2. Автомат ангилал ба бүдүүвч зураглалыг кодлох

Sorting automation and encoding schemes

Кодчлолын асуудалд эргэж орсны дараа автомат ангиллалд гарч буй асуудлууд нь бидний санаа зовж буй асуудал юм. Хэдийгээр яг тодорхой бүдүүвч зураглалыг кодлоход болон автомат ангиллалын хэрэгжүүлэлтийн талаар ярих боломжгүй ч, бүдүүвч зураглалыг кодлох болон автомат ангиллал нь үндсэндээ хоёр өөр хэмжүүртэй хэдийч хоорондоо огтлолцдог.
As we turn back to the encoding issue, problems that arise in sorting automation is what we are concerned about. Although it is impossible to talk of implementation of automated sorting without addressing a specific encoding scheme, encoding schemes and automated sorting are essentially two dimensions and intersect each other.

Авиа зүй болон дүрсээр кодчилох аргад зориулагдсанавтоматаар ангиллалын талаарх тайлбарыг 1-р хүснэгтэд үзүүлэв.Үсэгийг ангилахын тулд нөхцөл байдлаас үл хамаарсан кодлогдсон үсгийг ангилахад чиглэгддэг. Хэд хэдэн ялгах тэмдэглэгээг эс тооцвол (авиа зүйн кодчилолд FVS гэх мэт) ялгах тэмдэгтүүдийг үсгээр тэмдэглэсэн байдаг.

Remarks on various automated sorting methods for phonetic and graphetic approaches are summarized as in Table 1. Here character sorting refers to sorting encoded characters without any contextual sensitivity (expansion or contraction). Except for a few diacritical characters (such as FVSes in phonetic encoding), character sorting treats encoded characters as letters.

Хүснэгт 1. Автомат ангиллалын арга × кодлох аргачилал
Table 1. Automated sorting methods × encoding approaches

<table>
<tr><td rowspan="2"></td><td rowspan="2">Үсгийг ангилах
Character sorting</td><td colspan="2">Энгийн ангилал
Natural sorting</td></tr>
<tr><td>Ангилчлалыг задлах
Unfolded sorting</td><td>Ангилчлалыг нэгтгэх
Folded sorting</td></tr>
<tr><td>Авиа зүйн кодчилол
Phonetic encoding</td><td colspan="2">Найдваргүй
Unreliable</td><td>Хэрэгжиж болохуйц
Practicable</td></tr>
<tr><td>Дүрсийн кодчилол
Graphetic encoding</td><td>Ашиглах боломжгүй
Unusable</td><td>Боломжгүй
Impossible</td><td>Хэрэгжиж болохуйц
Practicable</td></tr>
</table>

Авиа зүйг кодлох тогтолцоо

For phonetic encoding schemes:

- Үсгийг ангилах болон ангиллалыг задлахад ихэнх тохиолдолд адил үр дүн гардаг. Өнөөгийн хамгийн өргөн хэрэглэгддэг Монгол хэлний автомат ангиллах аргын үр дүнд нь найдваргүйбайна. Учир нь нэг оруулж байгаа хувилбар нь өөр өөр үсгийг илэрхийлж хэрэглэгч тус бүр өөр өөр тэмдэгтийг ашиглаж байгаатай холбоотой.Энэхүү бичилтэнд нийцэхгүй байгаа хэд хэдэн шалтгаан бий. Үүнд:
  Character sorting and unfolded sorting yield identical results in most of the cases. Being the most widely used automated sorting of Mongolian today, however, the sorting results are unreliable, because one entry can be represented with different character sequences by different users. There are several reasons for this typing inconsistency:
  - Бичилтийн алдаа: Бичээч нар бичиж байх явцдаа алдаагаа анзаарахгүй байх.
    Typo: Typists may not always be able to notice their typing errors.
  - Үсгийг буруугаар ашиглах: Бичээч нар үсгийн хэрэглээ өөр байгааг тодорхойлж болох талтай.
    Misuse: Typists may identify letters differently.
  - Хэтрүүлэн ашигах: Бичээчид өөрсдийн зорилгодоо хүрэхийн тулд зарим буруу товчлууруудыг санаатайгаар ашигладаг. Тиймээс, янз бүрийн эх сурвалжаас бичсэн текстийг нэгтгэх үед нэг оруулгын ангилах үр дүн нь хэд хэдэн газруудад гарч ирдэг.
    Abuse: Typists may favor some incorrect keystrokes deliberately for the sake of expediency. Therefore, when text from different sources is aggregated, one entry can appear in several places of the sorting result.
- Нэгтгэх ангилал нь авиа зүйг кодчилоход хэрэгжих боломжтой, яагаад гэвэл энэ нь мөн л дүрсээр кодчилоход хэрэгжих боломжтой байдаг. (доорхийг уншина уу) Авиа зүйн кодчилол нь дүрсээр кодчилохын тулд дүрсээр кодчилох бүх мэдээлэлийг дамжуулах шаардлага гардаг.
  Folded sorting is practicable for phonetic encoding because it is also practicable for graphetic encoding (see below), and phonetic encoding contains all information graphetic encoding conveys and beyond.

Дүрсээр кодлох тогтолцоо

For graphetic encoding schemes:

- Шууд дүрсийн үсгийг ангилахдаа, хамгийн бага өртөгөөр хэл шинжлэлийн хувьд тодорхойгүй болон хэрэглэгчидэд ч мөн ойлгоход бэршээлтэй байдаг.
  Direct graphetic character sorting, being the least costly method, is not linguistically significant, and is incomprehensible to users.
- Ангиллах аргыг задлах боломжтой, учир нь кодлогдсон текстээс хойш адил бичигддэг өөр утгатай үг болон үсгийн хэлбэрийн хооронд ялгаа байдаг.
  Unfolded sorting is impossible, because the distinctions between homographic letterforms are missing from the encoded text.
- Дүрсийг кодлон ангиллах аргыг нэгтгэхэд хэрэгжүүлж болохуйц байна.
  Folded sorting of graphetic encoding is practicable.
  - Энэ нь хэрэглэгчдэд ойлгомжтой болж, ингэснээр 1.2-хэсэгт хэлэлцсэнзаасанчилан ашиглагддаг.
    It is comprehensible to users and thus usable, as has been discussed in Section 1.2.
  - Энэ нь техникийн хувьд хэрэгжих боломжтой, учир нь үсгийн нэгтэгсэн мэдээлэл нь нөхцөл байдлаас хамааран өөрчлөгдсөн дүрсийн холбооноос сэргээн босгогддог. Үүнийг доор тайлбарласан байгаа.

It is technically implementable, because the information of folded letters is still largely reconstructible from graphetic strings through a series of contextual transformations, which I will demonstrate shortly.

Автоматаар дүрсийг кодчилон ангилах, ялангуяа G2P ангилах арга нь маш их ач холбогдолтой юм. Учир нь энэ нь Монгол хэлийгкодчилсон монгол хэл дээрх автоматаар ангилах цорын ганц арга юм.
Automated folded sorting of graphetic encoding, or G2P sorting, is particularly of great significance because it seems to be the only practical solution to automated sorting for graphetically encoded Mongolian.

3. G2P ангилалын танилцуулага
Introduction to G2P sorting

G2P ангилал арга нь хоёр үе шаттайгаар хэрэгжсэн. Үүнд:
G2P sorting is implemented in two steps:

- 1-р үе: дүрсийн үсгийн холбоосоос үсгийг нэгтгэх дараалалыг дахин сэргээн завсарладаг.
  Step 1: Reconstruct folded letter sequences from graphetic character strings.
- 2-р үе: Оруулгуудыг үсэгийг нэгтгэх дарааллаар нь ангилдаг.
  Step 2: Sort the entries by their folded letter sequences.

G2P ангилах чадвар нь Юникодын Харьцуулалтын Алгоритмийн гол алгоритм дотор нөхцөл байдлаас хамаарсан хязгаарлагдмал нөхцөлөөс шалтгаалан багтаж чаддаггүй. (харьцуулах элементийн хүснэгтэд цөөн хэдэн товчлол нэмэхэд ажиллана, Thai/Lao CV гэх мэт хувиргаж оруулана гэсэн үг). Гэсэн хэдий ч, үсгийн дүрсийг үсгийн нэгтгэл уруу буулгахдаа хам сэдвээр нь болгоомжтой буулгана. Үүний үр дүнд, ангилах үндсэн процессоос өмнө дүрсийн нэгтгэх эгнээнд сэргээн засварлахад тусдаа өөрчлөлт хийх шаардлагатай болно.
G2P sorting cannot be accommodated within the main algorithm of the Unicode Collation Algorithm, as only limited contextual sensitivity (to the extent that adding a few contractions to the collation element table will work, like Thai/Lao CV inversion) can be handled thereby. However, the mapping from graphetic characters to folded character is highly context-sensitive. As a result, a separate step of transformation is needed to reconstruct folded strings before the main procedure of sorting.

Энэхүү хэлэлцүүлэг нь үсгийн нэгтгэх аргыг бүтээж сэргээн засварлах практик арга байгаа гэдгийг харуулах зорилготой юм. Гэсэн хэдий ч үндсэн дүрсээр кодчилох тогтолцоог үгийн дунд ордог хоёр шүд (ᠮ) Х, Г үсгийг төлөөлөх боломжтойэсвэл Г үсэг үгийн дунд орохдоо хоёр салангид (ᠷ) нэг нэг шүд хэлбэрээр бичигдэг. Дүрсээр кодчлолын аргын дотор бүх хувилбаруудыг багтаахын тулд зөвхөн хамгийн хувирашгүй кодчиллын тогтолцооны хэрэгжилтийг энд авч үзхэд уншигчид үүнийг бусадтай нь тохируулахад хялбар болдог. Тус хэлэлцүүлэг нь дүрсийн кодчиллын хамрах хүрээ ямар нэгэн тодорхой хувилбаруудыг дэмжиж байгаагаар тайлбарлахгүй гэдгийг онцлон тэмдэглэх нь зүйтэй.
The present discussion is aimed at showing that there are always practical solutions to the reconstruction of folded letters, however graphetically radical the encoding scheme is (as long as the medial double tooth (ᠮ) representing x or g is not broken into two single teeth (ᠷ)). In order to cover all practical variations within the graphetic encoding approach, only the implementation for the most radical graphetic encoding scheme is addressed here, and the readers should find it easy to adjust it to the rest. It should be emphasized that the present discussion shall not be construed as favoring any specific variant in the spectrum of graphetic encoding.

3.1 Ялгах элементийн хүснэгт

Collation element table

G2P-н ялгах элементийн хүснэгтийг Хүснэгт 2 харуулсан болно, Энд нэгтэгсэн үсэгүүдийг мөн тэдгээрийг ангилах дарааллаар нь жагсаасан байдал, харгалзах дүрсүүд, үсгийн дүрс, ангилах аргын нэгтгэсэн болон задаргасан байдлын дараалалуудыг тус тус харуулсан. Нэгтэгсэн үсэг, дүрсийн үсгийн хоёулангынх нь онцлогыг буулгахдаа жижиг үсгээр тэмдэглэж оруулсан. Харин дүрсийн үсгээр бичсэн тэмдэгтүүд нь том үсгээр бичигдсэн. Доорхи хүснэгтийг хархад, дүрс кодлох тогтолцооны үндсэн язгуурын чанараас хамааран 9 хүртэлх тоогоор илэрхийлэгддэг график дүрс байж болдог. Тухайлбал A, I, O, U, X, G, L, W, and H гэх мэт.

The collation element table of G2P sorting is given as Table 2, where folded letters are listed by their sorting order, and corresponding glyphs, graphic characters, and unfolded letters along with their orders are given together. Folded letters and graphetic characters in biunique mapping with them are transliterated with small letters, while graphetic characters not in biunique correspondence with folded letters are transliterated with capital letters. As can be seen from the table, there might be up to 9 under determined graphetic characters depending on the radicalness of the graphetic encoding scheme, namely A, I, O, U, X, G, L, W, and H. Secondary weights of collation elements in the table are largely arbitrarily given, as there is no widely accepted secondary weighting convention for either unfolded or folded sorting.

Хүснэгт 2. Ялгах элементийн хүснэгт

Unfolded order	Unfolded letter	Collation element	Folded letter	Graphetic character	Glyph			
					IS	I	M	F
—	—	[.00.3]	ʔ	A	□	ᠠ	ᠠ	□
1 2_1	*a e*	[.01.1]	*a*	A	□	ᠠ	ᠠ	ᠠ/ᠡ
		[.01.2]	*α* = α		ᠡ	■	■	■
2_2	*e*	[.02.1]	*ɇ*	—	■	■	·	ᠡ
		[.02.2]	*ë*	W	■	□	ᠸ	ᠸ
3	*i*	[.03.1]	*ïj*	II	□	□	ᠢᠢ	□
		[.03.2]	*i*	I	ᠢ	ᠢ	ᠢ	ᠢ/ᠶ
4 5 6_1 7_1	*o u ö₁ ü₁*	[.04.1]	*o*	O	■	ᠣ	ᠣ/ᠤ	ᠣ/ᠤ
		[.04.2]	*u*	U	ᠤ	■	■	ᠤ
6_2 7_2	*ö₂ ü₂*	[.05.1]	*ö*	OI	■	■	ᠥ/ᠦ	□
		[.05.2]	*ü* = ü		■	■	■	ᠥ/ᠦ
11	*n*	[.11.1]	*n* = n		■	ᠨ	ᠨ	ᠨ/ᠩ
		[.11.2]	*ń*	A	□	□	ᠠ	ᠩ
12	*ŋ*	[.12.1]	*ŋ*	AG	□	□	ᠭ	ᠩ
13	*b*	[.13.1]	*b* = b		■	ᠪ	ᠪ	ᠪ
14	*p*	[.14.1]	*p* = p		■	ᠫ	ᠫ	ᠫ
15_1	*x*	[.15.1]	*x*	X	■	ᠬ	ᠮ	ᠬ
16_1	*g*	[.16.1]	*ġ* = ġ		■	ᠭ	ᠭᠭ	ᠬ/ᠭ
		[.16.2]	*ǵ*	X	■	□	ᠮ	ᠬ
15_2 16_2	x_2 g_2	[.17.1]	*g*	G	■	ᠭ	ᠭ	ᠩ
17	*m*	[.18.1]	*m* = m		■	ᠮ	ᠮ	ᠮ
18	*l*	[.19.1]	*l*	L	■	ᠯ	ᠯ	ᠯ
19	*s*	[.20.1]	*s* = s		■	ᠰ	ᠰ	ᠰ
20	*š*	[.21.1]	*š* = š		■	ᠱ	ᠱ	ᠱ

$21_1\ 22_1$	$t_1\ d_1$	[.22.1]	t = t	■	ᠳ	ᠳ	ᠳ
$21_2\ 22_2$	$t_2\ d_2$	[.23.1]	d = d	■	ᠳ	ᠳ	ᠳ
		[.23.2]	$\eth$ OA	□	■	ᠳ	ᠳ᠊
23	$\check{c}$	[.24.1]	$\check{c}$ = č	■	ᠴ	ᠴ	ᠴ
24	$\check{j}$	[.25.1]	$\check{j}$ = ǰ	■	■	ᠵ	ᠵ
		[.25.2]	$\jmath$ I	᠈	᠈	□	□
25	y	[.26.1]	y = y	■	ᠶ	ᠶ	■
		[.26.2]	$\check{\imath}$ I	□	᠈	᠈	᠈
26	r	[.27.1]	r = r	■	ᠷ	ᠷ	ᠷ
27	w	[.28.1]	w W	■	ᠸ	ᠸ	ᠸ
		[.28.2]	$\check{u}$ U	□	■	■	ᠦ
28	f	[.29.1]	f = f	■	ᠹ	ᠹ	ᠹ
29	k	[.30.1]	k = k	■	ᠺ	ᠺ	ᠺ
30	c	[.31.1]	c = c	■	ᠼ	ᠼ	ᠼ
31	z	[.32.1]	z = z	■	ᠽ	ᠽ	ᠽ
32	h	[.33.1]	h H	■	□	ᠾ	ᠾ
		[.33.2]	$\hbar$ AH	■	ᠠᠾ	□	□
33	$\check{r}$	[.34.1]	$\check{r}$ = ř	■	ᠬ	■	■
34	$\dot{\imath}$	[.35.1]	$\dot{\imath}$ LH	■	ᠯᠾ	ᠯᠾ	□
35	$\hat{z}$	[.36.1]	$\hat{z}$ H	■	ᠽ	□	□
36	$\hat{c}$	[.37.1]	$\hat{c}$ OO	■	ᠶᠶ	□	■

3.2 **График дүрслэлээс нэгтэгсэн үсэгийг дахин сэргээх**

Reconstructing folded letters from graphetic characters

Аз болоход, энгийн орлуулалтын програмчилал нь(хэсэгчилэхгүй, давталтгүйгээр) бидний хүсэн тэмүүлж байгаа үсгийг дахин ангилахад илүү сайн хангаж өгч байна. Мөн толь бичгийн зүйлийг агуулаагүй үг хэллэгээр орлуулсан багц орлуулалтын талаар доорхи 3-х хүснэгтэд жагсааж бичсэн. Өөрчлөлтийг бүрэн дүүрэн болгох бүх орлуулалтыг тодруулсан болно. Бүлгүүдийг ((...)) эсвэл тоон дайвар үгийг (*, +,?, {m, n}, гэх мэт) энд эрэмбэлж оруулаагүйг дүрдах нь зүйтэй байх.

Luckily, finite unilinear (no branching, no looping) application of regular expression substitutions will suffice for our purposes of folded letter reconstruction. A set of substitutions containing no dictionary items for content words is adopted, as is listed in Table 3. Catch-all substitutions that ensure completeness of transformation are highlighted. It is worth mentioning that devices like groups ((…)) or quantifiers (*, +, ?, {m,n}, etc.) are not even resorted to here.

Хүснэгт 3: Нэгтэгсэн үсгийн сэргээн засварлахад хэрэглэдэг орлуулалтын жагсаалт

Table 3 List of regex substitutions employed in folded letter reconstruction

Search	Subs.
\bOOA\b	oð
\bI\b	i
\bIIAA\b	iĭań
\bIIAr\b	iĭar
\bOOI	ĉi
\bH	ẑ

\bAH	ɦ
\bLH	ɬ
\bW	w
\bI	ɟ
I(?=[ɑ])	ǐ
U(?=[ɑ])	ǔ
U	u
H	h
L	l
OI\b	oi
OII\b	öi
(?<=\bG)OII	öi
OII	oï̈
OI	ö
(?<=\BA)II\b	ǐi
(?<!I)I(?!I)	i
\bO	o
O(?!A)	o
(?<!A)G	g
(?<=\B[AWO])II	ï̈
I	i
OAG(?![AIOɑüaɐ̈ëiïjouö])	oŋ
(?<=\B[AIOɑüaɐ̈ëiïjouö])W	w
(?<=\B[AWIɑüaɐ̈ëiïjouö])OA	ð
(?<!A)G	g
O	o
X(?=[ǔAIɑüaɐ̈ëiïjouö])	x
X	ǵ
GW(?=[nbpmsštdčǰrfkczglɟwhġyř?xǐǔžĉɦɬ])	gë
G(?=[AIɑüaɐ̈ëiïjouö])	g
(?<=[nbpmsštdčǰrfkczglɟwhġyř?xǐǔžĉɦɬ])AAG	aŋ
WAAG	waŋ
(?<=[nbpmsštdčǰrfkczglɟwhġyř?xǐǔžĉɦɬ])AG	ag
(?<=[Iɑüaɐ̈ëiïjouö])AG	ŋ
WAG	ëŋ
(?<=[nbpmsštdčǰrfkczglɟwhġyř?xǐǔžĉɦɬ])AA	ań
WAA	wań
(?<=[nbpmsštdčǰrfkczglɟwhġyř?xǐǔžĉɦɬ])A	a
(?<=[Iɑüaɐ̈ëiïjouö])A	ń
(?<=\bA)WA(?![ńŋǵðɑüaɐ̈ëiïjouö])	ëń
(?<=[nbpmsštdčǰrfkczglɟwhġyř?xǐǔžĉɦɬ])WA(?![ńŋǵðɑüaɐ̈ëiïjouö])	ëń
WA	wa
AAAG	ʔaŋ
AAG	ʔɐ̈ŋ
AG	ʔɐ̈g
G	g
AAA	ʔań
AA	ʔa
A(?=[Wɑüaɐ̈ëiouö])	ʔ
A	ʔɐ̈
W(?=[i]\B)	w
W(?=[ïjńŋǵð])	ë

W(?=[u])	ë
(?<=[ʔɦɨ])W	ë
(?<=[Iɑüaɐ́ëiïjouöńŋǵð])W	w
WW	ëw
W	ë

4. G2P ангилалыг туршиж үзэх.
Testing G2P sorting

Өмнөх хэсэгт тодорхойлсон G2P ангилах аргыг 26433-үг зөв бичгийн өгөгдлийн сантай тулган шалгадаг. Үр дүн нь дараах байдалтай байна.
G2P sorting method specified in the previous section is tested against a 26433-word spelling database. The results are as follows.

4.1 Нэгтэгсэн үсгийг буруу сэргээсэн
Misreconstruction of folded letters

График дүрслэлээс нэгтэгсэн үсгийг дахин сэргээх мэдээлэл нь анхны задарсан үсэгтэй маш сайн нийцэж байгааг 2-р хүснэгтээс харж болох хэдий ч, 4-р хүснэгтэд үл хамаарах зүйлүүдийг жагсааж бичсэн.
The folded letters reconstructed from graphetically encoded data matched the original unfolded letters quite well in the correspondences defined in Table 2, except for the exceptions listed in Table 4 below.

Хүснэгт 4. Буруу сэргээгдсэн төрлүүд
Table 4 Types of misreconstruction

Төрлүүд Type	Задарсан үсэг Unfolded letter	Нэгтэгсэн үсэг Folded letter	Нэгтгэн дахин сэргээсэн үсэг Reconstructed folded letter	График үсэг Graphetic character	Ерөнхий үгийн гарал зүй Major etymological class	Тоо Count
A	e	ʔɐ́	ʔ	A	Уугуул Native	27/26433
B	Un	oń	ð	OA	Уугуул/Зээлсэн Native/Loan	12/26433
C	w ë	W ë	Ë w	W	Зээлсэн Loan	6/26433
D	Алдаатай зөв бичигдсэн Ill-formed spelling				Уугуул Native	2/26433

Хүснэгтээс хархад буруу сэргээгдсэн дөрвөн үндсэн төрөл байгааг харж болно: Үүнд:
As can be seen from the table, there are four major types of misreconstruction:

- Төрөл A: *ei* эсвэл *eü* үсгээр эхэлсэн бүх үг болон *en* эхэлсэн бүх үгнүүдийн дараа эгшиг дагаагүйг харж байна. Type A: All words beginning with ei or eü, and all words beginning with *en* which is not followed by a vowel.
- Төрөл B: Бүх үг on/un/ö n/ü n эдгээрийг багтаасан байдаг ба эдгээр нь эгшигийн өмнө орсон байх ба эгшгийн ард дагаж бичигддэггүй. Ихэнхдээ харьяалахын тийн ялгалд хамаарлтай хэдий ч, заримдаа гадаад үгнээс зээлж авсан үгэнд мөн харагддаг.
Type B: All words containing on/un/ö n/ü n which is preceded by a vowel but not followed by a vowel. Mostly in genitive stems of fugitive-n native words (e.g., EREÜ: nom. ereü, gen. ereü n=ü), but may also occur in loan words.

- Төрөл C: Цөөн хэдэн үгэнд *w* эсвэл *ë* орсон байдаг. Нөлөөлөлд өртсөн үгсийн жагсаалт сонгосон хөрвөх зарчмаас хамааран өөр өөр байж болдог.
  Type C: A few words containing w or ë. The list of affected words may vary, depending on the specific transformational rule set chosen.
- Type D: Маш цөөн хэдхэн үг алдаатай зөв бичигдсэн байдаг байна.
  A few words with ill-formed spelling.

Энэ буруу сэргээх гэдэг нь ерөнхийдөө урьдчилан таамаглах боломжтой гэсэн үг юм. 5-р хүснэгтэнд буруу сэргээгдсэн төрлүүдийн талаарх жишээнүүдийг харуулсан байгаа.
This means that cases of misreconstruction are predictable in general. Table 5 gives some examples of misreconstruction of each type.

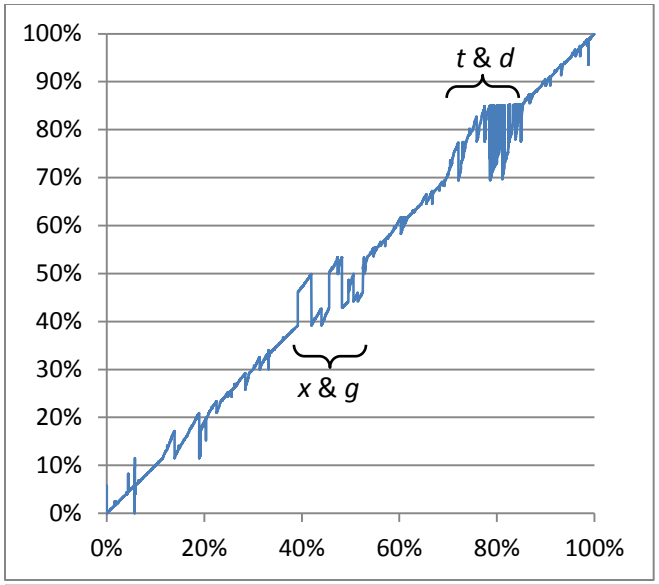
Хүснэгт 5.
Table 5

Төрлүүд Type	Задарсан үсэг Unfolded letter	Нэгтэгсэн үсэг Folded letter	Нэгттгэн дахин сэргээсэн үсэг Reconstructed folded letter	График үсэг Graphetic character
A	ei ein eü ende	ʔɐ̈i ʔɐ̈ijń ʔɐ̈u ʔɐ̈ńda	ʔi ʔiiń ʔu ʔada	AI AIIA AU AAdA
B	ereün ondoon	ʔɐ̈raoń ʔońdooń	ʔɐ̈rað ʔońdoð	ArAOA AOAdOOA
C	niswanis nirwalaxu burwasad yëlwi	niswanis nirwalaxu borwasað yëlwi	nisëńnis nirëńlaxu borëńsað yëlëi	nIsWAnIs nIrWALAXU bOrWAsAOA yWLWI
D	tŋri tŋrilig	tŋri tŋrilig	tagri tagrilig	tAGrI tAGrILIG

26433-үгийн мэдээллийн санд, зөвхөн 47 буруу сэргээлттэй оролтын тохиолдол байдаг хэдий ч нийт 0.17% -ийг эзэлж байна.
In the 26433-word database, only 47 entries of misreconstruction occur, making up 0.17% of the total.

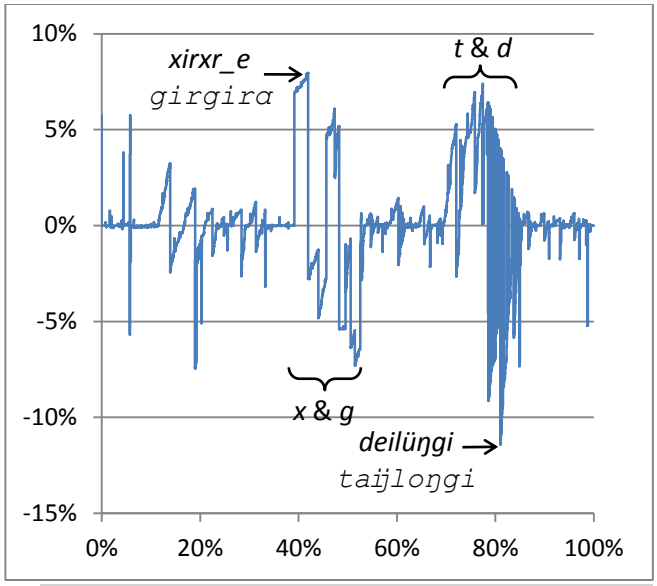
4.2 G2P ангилалыг хийсвэр задалдаг ангилалтай харьцуулах нь
Comparison of G2P sorting with ideal unfolded sorting

G2P ангилалын үр дүнг хийсвэр задалдаг ангилалтай харьцуулан хархад, Үр дүнгийн зураг 1 ба зураг 2-т харьцуулан харуулав. Үүнд:



Зураг 1: G2P ангилалын хийсвэрэрэр задалдаг ангилал

Figure 1 G2P sorting ~ ideal unfolded sorting



Зураг 2: G2P ангилалын чиглэл, сааталыг хийсвэр задалддаг ангилалтай харьцуулах

Figure 2 Lead and lags of G2P sorting relative to ideal unfolded sorting

Зураг 1 дээр G2P ангилалын дараалалыг хийсвэрээр задалдаг ангилалтай хувиар харуулав. Зураг 2 дээр ялгаануудыг харьцуулан харуулав. G2P-ийн хамгийн дээд эрэмбийн ялгаа нь -11.41% -иас ялимгүй ялгаатай, *deilüŋgi*-н хувьд ангилахдаа ***taijloŋgi*** хэлбэрээр хамгийн их хоцролт + 7.96%ангилсан байгааг харж байна. Харин ***xirxir_e***-н хувьд ангилахдаа ***girgira*** хэлбэрээр ангилсан байна. Бүх 26433 оруулгуудын ялгах дарааллын стандарт хазайлт (σ) нь 2.84% байдаг.

Figure 1 shows the relative order of G2P sorting against ideal unfolded sorting in percentage, and Figure 2 shows their difference. The maximum lead of G2P sorting to ideal unfolded sorting is −11.41%, occurring at ***deilüŋgi*** sorted as t***aijloŋgi***, while the maximum lag is +7.96%, occurring at ***xirxir_e*** sorted as ***girgira***. The standard deviation (σ) of the sorting order differences of all 26433 entries is 2.84%.

Ангилах ялгааны гол хувь нэмэр нь эхний үсэгний x / g ба t / d хосуудаас бүрдэх ба эхний эгшигний / e, o / u ба ö / ü хосуудаас ихээхэн хамаардаг. Үсгийг задлан ангилах аргачилалд дассан хэрэглэгч нь G2P-ыг ямар нэгэн сургалтгүйгээр дасан зохицоход хялбар байдаг, зөвхөн энэ орооцолдсон хос бичсэн хэлбэрээрээ тохируулахыг санаарай.

The major contributions to the sorting differences come from x/g and t/d pairs of the first letter, and less significantly from a/e, o/u and ö /ü pairs of the first vowel. A user accustomed to ideal unfolded sorting should find it easy to adapt to G2P sorting without much training, just to remember to make some adjustment to these tangled pairs according to their written shapes.

Адил төстэй байдлыг хялбархан ангилиах үүднээс ялгах аргын үр дүнг 6-р хүснэгтэд үзүүлэв. G2P ангилалын мэдээлэлийн санд G2 үсэг задарч байгаа тэр хэсэгт жигд хэлбэрээр явагддаг. Нийт 52 өөр оруулгаас зөвхөн 5 (тодруулсан) оруулга боломжтой гэсэн үр дүн гарч байна. Эцэст нь хэлхэд, G2P ангилал нь ямар нэгэн байдлаар хазаахгүй хамгийн тохиромжтой нь гэж үзсэж болно

To give an intuitive view of the sorting similarity, an extract of the results of both sorting methods are given in Table 6, where the items are extracted in a uniformly spaced manner from the G2P-sorted database. The results show that only 5 entries (highlighted) out of total 52 differ. In conclusion, G2P sorting does not deviate as much from ideal unfolded sorting as one might expect.

6-р хүснэгт: G2P ангилалын задаргаа

Table 6 Extract of G2P sorting and ideal unfolded sorting

Rel. delta sort. ord.	Unfolded sort. ord	Unfolded letter	G2P sort. ord.	Reconstructed folded letter	Graphetic character
0.06%	517	agawa	500	ʔaġawa	AAġAWA
0.00%	1000	asagtuxu	1000	ʔasaǵdoxu	AAsAXdOXU
−0.05%	1488	arčigur	1500	ʔarčiġor	AArčIġOr
0.06%	2015	elegdexü	2000	ʔɐlagdago	ALAGdAGO
0.04%	2511	iin	2500	ʔiiń	AIIA
0.01%	3003	irgagalaxu	3000	ʔirġaġalaxu	AIrġAġALAXU
−0.91%	3259	obosxixü	3500	ʔobosgigo	AObOsGIGO
0.62%	4164	usadxagči	4000	ʔosaðxaǵči	AOsAOAXAXčI
−0.14%	4462	ursigtai	4500	ʔorsiǵdai	AOrsIXdAI
	(5000	örgedxel)			
1.04%	5276	ülemǰidexü	5000	ʔölamǰidago	AOILAmǰIdAGO
−1.89%	5000	örgedxel	5500	ʔörgaðgal	AOIrGAOAGAL
−0.63%	5834	namči	6000	namči	nAmčI
−0.09%	6475	noxaituxu	6500	noxaijdoxu	nOXAIIdOXU
−0.01%	6997	bagaturčud	7000	baġadorčoð	bAġAdOrčOOA
0.71%	7689	beyeči	7500	bayači	bAyAčI
1.23%	8326	buurǰi	8000	boorǰi	bOOrǰI
0.25%	8565	buǰigirtuxu	8500	boǰigirdoxu	bOǰIGIrdOXU
0.02%	9005	bürxüixü	9000	börgoijgo	bOIrGOIIGO
0.00%	9500	xabursil	9500	xaborsil	XAbOrsIL
0.00%	10000	xasumal	10000	xasomal	XAsOmAL
2.56%	11179	xoordalg_a	10500	xoordalġɑ	XOOrdALġɑ
1.82%	11482	xošoŋčilaxu	11000	xošoŋčilaxu	XOšOAGčILAXU
	(12250	xög)			
5.39%	12928	gagaglaxu	11500	ġaġaǵlaxu	ġAġAXLAXU
6.75%	13788	gutugaxu	12000	ġodoġaxu	ġOdOġAXU
2.63%	13198	gemsixü	12500	gamsigo	GAmsIGO
1.20%	13319	gilalǰam_a	13000	gilalǰamɑ	GILALǰAmɑ
	(13788	gutugaxu)			
−4.72%	12250	xög	13500	gög	GOIG
0.45%	14120	güyüldüxü	14000	göyoldogo	GOIyOLdOGO
0.34%	14590	mesil	14500	masil	mAsIL
−0.01%	14997	möŋxelexü	15000	möŋgalago	mOIAGGALAGO
−0.11%	15471	sanagatai	15500	sanaġadai	sAnAġAdAI
−0.80%	15789	salbarxai	16000	salbarxai	sALbArXAI
0.05%	16514	sibeg	16500	sibag	sIbAG
−0.03%	16993	siratuxu	17000	siradoxu	sIrAdOXU
−0.82%	17283	soyoxai	17500	soyoxai	sOyOXAI
−0.02%	17995	šaliyatuxu	18000	šaliyadoxu	šALIyAdOXU
−0.23	18440	taitagar	18500	taijdaġar	tAIIdAġAr
0.86%	19227	tebxe	19000	tabga	tAbGA
6.16%	21133	dagir	19500	tagir	tAGIr
−4.21%	18884	tasixu	20000	tasixu	tAsIXU

−5.30%	19097	tawarčilaxu	20500	tawarčilaxu	tAWArčILAXU
	(19227	tebxe)			
−4.68%	19761	togonočaxu	21000	toġonočaxu	tOġOnOčAXU
−4.62%	20275	tulgagurtai	21500	tolġaġordai	tOLġAġOrdAI
	(21133	dagir)			
2.03%	22537	düŋsüixü	22000	töŋsoijgo	tOIAGsOIIGO
0.39%	22604	dürsütei	22500	törsodai	tOIrsOdAI
−0.48%	22874	časuraxu	23000	časoraxu	čAsOrAXU
0.00%	23500	čisurxau	23500	čisorxau	čIsOrXAU
0.35%	24092	čüüreǰexü	24000	čöoraǰago	čOIOrALǰAGO
0.87%	24730	ǰegülte	24500	¸jagolda	IAGOLdA
−0.01%	24998	ǰilabči	25000	¸jilabči	IILAbčI
−0.46%	25378	ǰoltai	25500	¸joldai	IOLdAI
0.43%	26114	yexemsüg	26000	yagamsog	yAGAmsOG

5 ДҮГНЭЛТ

5 Summary

Тус баримт бичигт тусгагдсан гол чухал асуудлууд нь:

The key points of this document are:

- Нэгтгэн ангилах аргачилал нь шинэ нээлт биш, мөн энэ нь задлан ангилалд муу үр дагавар үзүүлдэггүй.
  Folded sorting is not a new invention, nor is it inferior to unfolded sorting.
- G2P ангилах аргачилал, нэгтгэн ангилах хувилбар болох нь графикаар кодчилох шаардлагатай байгаа Монгол хэллэгийг автоматаар ангилах хамгийн сайн шийдэл байж болох юм.
  G2P sorting, as a variant of folded sorting, can be a good solution when automated sorting of graphetically encoded Mongolian is needed.
- G2P-ийн ялгах ангилалын үр дүн нь өргөн хэрэглэгддэг задаргаатай ангиллуудын хувьд ялгаатай биш бөгөөд хэрэглэгчид үүнийг хялбархан ашиглах боломжтой юм.
  The results of G2P sorting are not dissimilar to those of widely used unfolded sorting, and users will readily get used to it.

A. Толь бичгийн толгойн дарааллыг эрэмбэлэх аргачилал болон G2P ангилах аргачилал

A. Ordering of dictionary headings in ideal unfolded sorting and G2P sorting

7-р хүснэгтэд орчин үеийн Хятад дахь хамгийн өргөн дэлгэрсэн хувилбар болох толь бичгийн гарчигыг нэлээд задлан ялгаж харуулав. Доорхи хүснэгтэнд байгаа нэг эгнээнд онцлон тэмдэглэсэн үсэгнүүд нь адил бичигддэг өөр утгатай үсэгнүүд юм.

Table 7 shows an ideal unfolded sorting of dictionary headings, which is the most widely accepted version in modern China. Headings in adjacent cells with the same underline patterns are homographic.

Хүснэгт 7 Толь бичгийн гарчигийг ангилах

Table 7 Ideal unfolded sorting of dictionary headings

	a	$<_1$	*e* $<_2$ *ë*	$<_1$	*i*	$<_1$	*o*	$<_1$	*u*	$<_1$	*ö*	$<_1$	*ü*
$<_1$	*na*	$<_1$	*ne* $<_2$ *në*	$<_1$	*ni*	$<_1$	*no*	$<_1$	*nu*	$<_1$	*nö*	$<_1$	*nü*
$<_1$	*ba*	$<_1$	*be* $<_2$ *bë*	$<_1$	*bi*	$<_1$	*bo*	$<_1$	*bu*	$<_1$	*bö*	$<_1$	*bü*
						⋮							
$<_1$	*xa*	$<_1$	*xe* $<_2$ *xë*	$<_1$	*xi*	$<_1$	*xo*	$<_1$	*xu*	$<_1$	*xö*	$<_1$	*xü*
$<_1$	*ga*	$<_1$	*ge* $<_2$ *gë*	$<_1$	*gi*	$<_1$	*go*	$<_1$	*gu*	$<_1$	*gö*	$<_1$	*gü*
						⋮							
$<_1$	*ta*	$<_1$	*te* $<_2$ *të*	$<_1$	*ti*	$<_1$	*to*	$<_1$	*tu*	$<_1$	*tö*	$<_1$	*tü*
$<_1$	*d'a* $<_2$ *da*	$<_1$	*d'e* $<_2$ *d'ë* $<_2$ *de* $<_2$ *dë*	$<_1$	*d'i* $<_2$ *di*	$<_1$	*d'o* $<_2$ *do*	$<_1$	*d'u* $<_2$ *du*	$<_1$	*d'ö* $<_2$ *dö*	$<_1$	*d'ü* $<_2$ *dü*
						⋮							

8-р хүснэгтэд G2P ангилалын толь бичгийн гарчигийг үзүүлэв. "≡" -н хоёр тал дээрх гарчиг нь адил бичигддэг өөр утгатай үсэг. “⊃” гэдэг тэмдэглэгээ нь нэмэлт таамаглалыг илэрхийлнэ. Бага хаалтад байгаа тэмдэглэгээнүүд хийсвэр нэгтэгсэн ангилалыг илэрхийлнэ. Жнь: ***en*** гэдэг тэмдэг G2P ангилалын оронд хийсвэр нэгтэгсэн ангилалын ***e*** тэмдэглэгээний дотор гарчиглан ангилагдаж орно.

Table 8 shows G2P sorting of dictionary headings. Headings on both sides of “≡” s are homographic. “⊃”s indicate additional subsumptions, and the parenthesized terms should be in their normal places in ideal folded sorting. For example, ***en*** should be sorted under the heading of ***e*** in ideal folded sorting, instead of under a in G2P sorting.

Хүснэгт 8: G2P ангилалын толь бичигийн гарчиг аргачилал

Table 8 G2P sorting of dictionary headings

	a (⊃ *en*)	$<_1$	*e* $<_2$ *ë*	$<_1$	*i* (⊃ *ei*)	$<_1$	*o* ≡ *u* (⊃ *üi*)	$<_1$	*ö* ≡ *ü*
$<_1$	*na* ≡ *ne*	$<_1$	*në*	$<_1$	*ni*	$<_1$	*no* ≡ *nu* (⊃ *nüi*)	$<_1$	*nö* ≡ *nü*
$<_1$	*ba* ≡ *be*	$<_1$	*bë*	$<_1$	*bi*	$<_1$	*bo* ≡ *bu* (⊃ *büi*)	$<_1$	*bö* ≡ *bü*
					⋮				
$<_1$	*xa*					$<_1$	*xo* ≡ *xu*		
$<_1$	*ga*					$<_1$	*go* ≡ *gu*		
$<_1$	*xe* ≡ *ge*	$<_1$	*gë*	$<_1$	*xi* ≡ *gi*			$<_1$	*xö* ≡ *xü* ≡ *gö* ≡ *gü*
					⋮				
$<_1$	*ta* ≡ *te* ≡ *da* ≡ *de*	$<_1$	*të* ≡ *dë*	$<_1$	*ti* ≡ *di*	$<_1$	*to* ≡ *tu* (⊃ *tüi*) ≡ *do* ≡ *du* (⊃ *düi*)	$<_1$	*tö* ≡ *tü* ≡ *dö* ≡ *dü*
$<_1$	*d'a* ≡ *d'e*	$<_1$	*d'ë*	$<_1$	*d'i*	$<_1$	*d'o* ≡ *d'u* (⊃ *d'üi*)	$<_1$	*d'ö* ≡ *d'ü*
					⋮				

Толь бичгийн гарчигийг Худам бичгийн задарсан болон нэгтэгсэн үсэгний аль алинаар жагсаасан хүснэгтийг энд тайлбар өгч ойлгомжтой байхын тулд харууллаа .

Tables of dictionary headings in Hudum listed by both unfolded and folded letters are given here for the convenience of reference.

Хүснэгт 9:Худам толь бичгийн гарчиг

Table 9 Dictionary headings in Hudum

(а) Задарсан үсгээр жагсаасан хэлбэр

(a) Listed by unfolded letters

	-a	*-e*	*-ë*	*-i*	*-o/u*	*-ö/ü*
-						
n-						
b-						
⋮				⋮		
x-						
g-						
⋮				⋮		
t/d-						
d'-						
⋮				⋮		

(б) Нэгтэгсэн үсгээр жагсаасан хэлбэр

(b) Listed by folded letters

	−a	*−ȩ*	*−ë*	*−i*	*−o*	*−ö*
ʔ−						
n−		■				
b−		■				
⋮				⋮		
x−		■	■	■		■
ġ−		■	■	■		■
g−		■			■	
⋮				⋮		
t−		■				
d−		■				
⋮				⋮		

Б. График үсэг болон нэгтэгсэн үсгийн хоорондын уялдаа холбоо

B. Correspondence between graphetic characters and folded letters

График болон нэгтэгсэн үсгүүдийн хоорондын уялдаа холбоог ойлгоход хялбар болгох зорилгоор Хүснэгт 10 оруулаа . Нэг үсэг болон үсгийн нийлбэрүүдийг хүснэгтэд эгнүүлж бичиж оруулсан байна.

Correspondence between graphetic characters and folded letters is given as Table 10 for the convenience of reference. Single characters and character combinations are aligned in the table.

Хүснэгт 10: График болон нэгтэгсэн үсгийн хоорондын учлдаа холбоо

Table 10 Correspondence between graphetic characters and folded letter

Glyph				Graphetic character	Folded letter	Unfolded letter	Glyph				Graphetic character	Folded letter	Unfolded letter
IS	I	M	F				IS	I	M	F			
ᠠ	ᠠ	ᠠ	ᠠ/ᠨ	**A**	ʔ ʔ◌́ a ń	— e a n	□	□	ᠮ	ᠮ	**A**G	ŋ	ŋ
							■	ᠠ	□	□	**A**H	ɦ	h
							□	■	ᠳ	ᠳ/ᠽ	O**A**	ð	d
ᠢ	ᠢ	ᠢ	ᠢ/ᠶ	**I**	i j̱ ǐ	i j y	□	□	ᠬ	□	**II**	ij̱	i
							■	■	ᠥ/ᠦ	□	O**I**	ö	ö/ü
■	ᠣ	ᠣ/ᠤ	ᠣ/ᠤ	**O**	o	o/u/ö/ü	□	■	ᠳ	ᠳ/ᠽ	**O**A	ð	d
							■	■	ᠥ/ᠦ	□	**O**I	ö	ö/ü
							■	ᠪᠠ	□	■	**OO**	ĉ	ĉ
ᠤ	■	■	ᠤ	**U**	u ŭ	o/u/ö/ü w							
■	ᠬ	ᠬ	ᠽ	**X**	x ǵ	x g							
■	ᠭ	ᠭ	ᠨ	**G**	g	x/g	□	□	ᠮ	ᠮ	A**G**	ŋ	ŋ
■	ᠯ	ᠯ	ᠯ	**L**	l	l	■	ᠯᠠ	ᠯᠠ	□	**L**H	ɬ	ɬ
■	ᠸ	ᠸ	ᠸ	**W**	ë w	e w							
■	ᠰ	ᠰ	ᠰ	**H**	h ẑ	h ẑ	■	ᠠ	□	□	A**H**	ɦ	h
							■	ᠯᠠ	ᠯᠠ	□	L**H**	ɬ	ɬ