Egy vakmerő digitális lexikográfiai kísérlet: a CHDICT nyílt kínai-magyar szótár

Ugray Gábor

https://chdict.zydeo.net/
zydeodict@gmail.com

Kivonat: A CHDICT-tel egy nyílt, közösségileg szerkesztett kínai-magyar szótár indul hamarosan útjára. A cikk az előképek kontextusába helyezve mutatja be a munkát, illetve beszámol a kiinduló lexikai tartalom kiválasztásáról, az alkalmazott szótárfordítási eljárásról és közzététel módjáról. A szerző kísérletként tekint a projektre, melyből kiderülhet, származtatható-e kis nyelvpárokra kielégítő minőségű új szótár az elérhető nyílt forrásokból, és élteképes-e a kollaboratív modell ilyen tartalmakra.

Kulcsszavak: kínai-magyar, szótár, fordítás, nyílt, kollaboratív

1 Bevezetés

A digitális korban a nyelvi közösségek kulturális és gazdasági kibontakozását erősen befolyásolja, hogy mennyi és milyen minőségű digitális nyelvi erőforrást képesek előállítani és karbantartani. Jelen munka is ebben a tágabb összefüggésben értelmezhető: az alábbiakban egy nyílt, az interneten ingyenesen kereshető és bárki által szerkeszthető kétnyelvű szótár létrehozásáról számolok be.

Míg az említett kulturális javak előállítása komoly szellemi munkát igényel és a végtermék forintosítható értéket képvisel, ideális esetben legalább a nyelvi infrastruktúra alapelemeinek mindenki számára egyszerűen és ingyenesen hozzáférhetőnek kell lenniük. Erre ígérnek kézenfekvő megoldást a közösségileg szerkesztett, nyílt tartalmak.

A CHDICT hozzávetőleg 10 ezer szócikkel indul útjára a következő hónapokban, és a szigorú értelemben vett nyelvi tartalmon túl célja egy olyan interaktív online felület kialakítása, amely egyéb nyilvánosan hozzáférhető adattárak beépítésével minőségi javulást jelent a nyomtatott szótárakhoz képest.

A CHDICT létrehozásának csupán két tényező állt útjában. Az első, hogy mindeddig nem létezett ingyenesen felhasználható és szabadon hozzáférhető kínai-magyar szótár, amely kiindulási alapként szolgálhatott volna. A második, hogy a szerző valójában nem tud kínaiul. Utóbbi tényből ered a vállalkozás vakmerő jellege.

2 Előképek

A CHDICT mint nyílt, közösségileg szerkesztett tudástár nem előzmények nélküli. Kézenfekvő a Wikipediára gondolni, de ennél speciálisabb, szorosabban a tárgyhoz kötődő előképekre is támaszkodhattam.

Az első ilyen jellegű kezdeményezés az EDICT[1] japán-angol szótár [2] volt, amely 1991-ben indult útjára, és máig is fejlődik, immár 170 ezernél is több címszót tartalmaz. Az EDICT példáját követve az elmúlt 25 évben számos további szótár is létrejött, amelyek egy-egy kelet-ázsiai nyelvet (japánt vagy kínait) kötnek össze európai nyelvekkel. A teljesség igénye nélkül: 1997 óta fejlődik a kínai-angol CEDICT[2] (114 ezer címszó); 1999 óta a japán-német Wadoku[3] [7] (115 ezer címszó); 2006 óta a kínai-német HanDeDict[4] (150 ezer címszó). Magyar viszonylatban említést érdemel a HunNor norvég-magyar szótár,[5] amely egy maroknyi ember kezdeményezéseként indult a 2000-es évek elején, s mára 40 ezer szócikket tartalmaz.

Több mint két évtized tapasztalatai alapján leszűrhetünk néhány tanulságot ezekből a projektekből. Legtöbbjüknek sikerült elérni, sőt jócskán meghaladni a közepes méretet. Egyikük sem alkalmaz bonyolult formátumot a lexikai tartalom reprezentálására: mind a CEDICT, mind a HanDeDict máig az eredeti EDICT-formátumot követi, amelyben egy sor egy szócikknek felel meg.

A CEDICT története rámutat az átgondolt és explicit licencfeltételek fontosságára. Amikor a szerzői jogot implicit birtokló üzemeltető 2007-ben elérhetetlenné vált, a bizonytalanság az anyag továbbélését is veszélyeztette.[6]

A lenyűgöző „külső” terjedelem, azaz a szócikkek magas száma mellett megfigyelhető, hogy „befelé” a nyílt szótárak nem túl kiterjedtek: kevés nyelvtani és metainformációt közölnek, s a szélsőségesen egyszerű formátum miatt azt is kevéssé normalizált formában teszik. Feltételezésem szerint ez a közösségi szerkesztés velejárója: a legtöbb közreműködő nem rendelkezik nyelvészeti háttérrel.

3 Lexikográfiai eljárás

A nyílt kínai-magyar szótár elindításához először is egy tyúk-tojás problémát kellett feloldanom. Ha nincsenek közreműködők, nincsen közösségileg szerkesztett szótár sem. Ha azonban nincs olyan kiinduló anyag, amely méreténél fogva már értéket képvisel a felhasználók számára, nem lesz közösség sem, ami a szótárat bővítené és továbbfejlesztené.

Az alábbiakban leírt eljárás célja, hogy a rendelkezésre álló források maximális kihasználásával belátható időn belül elérjem a szükséges kiinduló állapotot, méghozzá

[1] http://www.edrdg.org/jmdict/edict_doc.html
[2] https://cc-cedict.org/wiki/
[3] https://www.wadoku.de/
[4] https://handedict.zydeo.net/
[5] http://dict.hunnor.net/
[6] Korabeli, immár nem fellelhető levelezőlisták tartalma alapján.

egész egyszerűen a kiválasztott címszavak CEDICT- és HanDeDict-beli angol és német megfelelőinek magyarra fordításával. A cél kimondottan egy *tökéletlen*, de „elég jó” szótár, ami az évek során szervesen javul és bővül majd.

3.1 Terjedelem

A terjedelem meghatározásához nem indulhattam ki az érett mintaképek méretéből. Az egyik kézenfekvő támpont a Kínai Népköztársaság hivatalos nyelvi szintfelmérőjéhez, a 汉语水平考试-höz (Hànyǔ Shuǐpíng Kǎoshì, HSK) közzétett szólista[7] volt. A legmagasabb szint teljesítéséhez elvárt szókincs 6 ezer szót tesz ki.

E lista tekintetbe vétele mellett szól, hogy a szárazföldi Kínába igyekvő nyelvtanulók mindenképpen a fenti vizsgára készülnek fel, így joggal várják el szótáruktól, hogy tartalmazza az előírt lexikai elemeket. Másrészt okkal lehetnek kétségeink, hogy a lista mennyire felel meg a kortárs nyelvhasználatnak. Számos jel mutat arra, hogy a legkorszerűbb tanulói szótárak kivételével a nyelvi segédanyagok többsége nincs szinkronban a tényleges modern nyelvhasználattal. A német nyelv esetén Tschirner [6] mutatta ki, hogy a 4 ezer szócikk nagyságrendű tanulói szótárak a korpuszokból okadatolható leggyakoribb szavak jelentős hányadát nem tartalmazzák, ellenben számos ritkább lexikai elemet feltüntetnek.

Az empirikus szógyakoriságokat a SUBTLEX-CH korpuszból [3] merítettem. A korpusz kínai filmfeliratokat foglal magában, azaz a kortárs köznyelvről ad képet. A közzétett gyakoriségi listát az teszi különösen értékessé, hogy valóban szavakat tartalmaz, nem írásjegyeket, ami a szóhatárokat nem jelölő kínai írás miatt ritkaságnak számít.

Másik írás tárgyát fogja képezni annak elemzése, hogy mennyire tekinthetjük relevánsnak és teljesnek a HSK-szólistát a SUBTLEX-CH korpusz gyakoriságainak tükrében.

Kézenfekvő viszonyítási pont volt a 10 ezres cél kitűzéséhez Bartos Huba és Hamar Imre kiváló, nyomtatott kínai-magyar szótára [1], amely a kiadó közlése szerint összesen 11.750 bejegyzést tartalmaz. Utólagos megerősítésként találtam Naszódi Mátyás megjegyzésére, miszerint „egy ember belátható idő alatt maximum 10.000 tételből álló szótár készítésére képes”. [5]

A CHDICT kiinduló törzse tehát a HSK-vizsgák 6 ezer szavát tartalmazza, kiegészítve a 4 ezer leggyakoribb szóval, amelyek a vizsgák anyagában nem szerepelnek.

3.2 Források

Eljárásom lényege, hogy a kiválasztott szócikkeket a CEDICT angol, illetve a HanDeDict német megfelelőiből magyarra fordítom. Ennek szerzői jogi szempontból nincs akadálya, mivel a Creative Commons licenc forrásmegjelölés mellett engedélyezi a származtatott anyagok létrehozását.

[7] http://www.hskhsk.com/word-lists.html

A CHDICT kezdeti minőségét a fenti két forrás korlátozza alulról, súlyosbítva a fordításból óhatatlanul adódó torzításokkal. A CEDICT-en és a HanDeDict-en kívül ezért kettős céllal több más forrást is tekintetbe veszek a munka során.

Az első cél a minőség javítása. Az angol jelentések fordításakor a fő problémát az angol nyelv szófaji és szemantikai többértelműsége jelenti. Ezt sajnos csak részben ellensúlyozza a HanDeDict bevonása, mivel ennek sok szócikkét eleve a CEDICT fordításaként állították elő. Egyéb források tekintetbe vételével az angolból eredő többértelműséget igyekszem ellensúlyozni.

A második cél a sebesség. A szótárfordításhoz dedikált eszközt fejlesztettem ki, amely a Google és a Bing fordítómotorokból származó gépi fordítások alapján gépelésgyorsító funkciókat nyújt az emberi fordítás bevitele során.

A szótárfordító alkalmazás a CEDICT-en és a HanDeDict-en túl tartalmazza még a Wikipediából származó cikkek címeit, ha a címszó szerepel önálló cikként, és ahhoz angol, német vagy magyar cikk is társul. Ehhez az előkészítési fázisban a Wikipedia letölthető adatbázis-mentéseit[8] dolgoztam fel gépileg.

Az előkészítés eredménye egy többnyelvű XML-fájl, amely címszavanként tartalmazza az összes eddig ismertetett forrást és azok gépi fordításait.

Másodlagos forrásként támaszkodok az ABC Chinese-English Dictionary-re [4], valamint a Bartos-Hamar-féle kínai-magyar szótárra. Ezeket szerzői jogi okokból nem használom fel módszeresen, de elszigetelt esetekben nagy segítséget jelentenek egyes szavak jelentésének tisztázásában. A MOEDICT kínai értelmező szótárat,[9] amely a tajvani sztenderd mandarint írja le, elsősorban a hagyományos írásjegyekkel és tajvani kiejtéssel kapcsolatos inkonzisztenciák feloldására használom. Avantgárd online „kutatási módszerként" említést érdemel még a Google képkeresési funkciója. Az elegendő szkepszissel kezelt eredmények időnként meglepő információkkal szolgálnak egy-egy szó regiszteréről, képzettársításairól, szerencsés esetben referenséről.

3.3 Munkakörnyezet

A szótárfordításhoz munkaeszközként először egy „polcról levehető", kereskedelmi fordítási környezetet vettem fontolóra. Hamar nyilvánvalóvá vált azonban, hogy itt a hagyományos fordítástól igen eltérő feladatról van szó, és érdemes kifejleszteni egy dedikált, egyszer használatos alkalmazást, amely a munka során a fejlesztési idő többszörösét takarítja meg.

[8] https://en.wikipedia.org/wiki/Wikipedia:Database_download
[9] https://www.moedict.tw/about.html

1. ábra. Képernyőkép a szótárfordításhoz használt munkaeszközről.

A felület explicit elemei elsősorban a hatékonyságot és gyorsaságot szolgálják. Ide tartozik az automatikus kiegészítési funkció a gépi fordításokból kigyűjtött szavak alapján, de a források gyorsan áttekinthető, tipográfiailag tagolt megjelenítése és a könyen navigálható címszólista is.

Implicit összetevő a címszavak sorrendezése. A lista alfabetikus rendezése óhatatlanul monotóniához vezetett volna, de nem előnyös a nyers gyakoriság alapú rendezés sem, mert egész másfajta kihívást jelentenek a gyakori, rövid és rendkívül poliszém kínai szavak, mint a ritkább, hosszabb és egyértelműbb elemek. Túl sok hasonló fejtörő egymás után szintén kedélyrombolú hatású.

Az optimális választás egy kétszintű rendezés volt. Az első rendezési szempont a gyakori szavakat sorolja felülre, viszont minden szó alá besorolja azokat a ritkább lexikai elemeket, amelyeknek a szó prefixe, vagy amelyek a szónak prefixei. Így a listán periodikusan váltakoznak a nagy és kis frekvenciájú szavak, s egymás közelébe kerülnek a jelentésükben összefüggő összetett lexikai elemek is. Mivel pedig a lista végén is szerepelnek gyakori szavak, nem lehetséges a lelkesedés alábbhagytával úgy dönteni, hogy mégiscsak elég lesz 7.500 szó, hiszen a hátralevő anyagban elszórva még rengeteg olyan elem szerepel, amelyről nem mondhatok le.

Az írás pillanatában 7.800 szócikk fordítása áll készen, s a tapasztalatok alapján 100 szócikk feldolgozása körülbelül két munkaórát vesz igénybe. Az eszköz minden egyes szócikkre rögzíti a munkaidőt, így később megvizsgálható, hogy van-e összefüggés az egyes szavak lefordításához szükséges idő és a frekvencia, a szó jelentésszáma, a fordítás minősége stb. között.

4 Az elkészült kiindulópont

4.1 Formátum

A szótár technikailag nem más, mint egy letölthető és szabadon felhasználható szövegfájl. Úgy döntöttem, hogy nem definiálok saját formátumot, vagyis a CHDICT is azt a formátumot használja majd, mint az EDICT, a CEDICT és a HanDeDict.

A változtatás mellett szólt volna, hogy ez az igen egyszerű formátum nem képes jólformáltsági feltételeket biztosítani a szófaji megjelölések, címkézett stílusjegyek, nyelvtani információk (pl. számlálószavak, többszótagos igék belső szerkezete, vonzatok) leírására, illetve nem ad lehetőséget az alternatív kiejtési változatok és írásmódok elegáns jelölésére sem.

Erősebbnek éreztem viszont a hátrányoknál azt a sokatmondó tényt, hogy három különböző, immár a százezres méretet meghaladó, közkedvelt szótár is remekül elboldogul a fenti korlátokkal. A bevett formátum további előnye, hogy megkönnyíti a szótár beépítését az elterjedt offline alkalmazásokba, mint a Pleco vagy a Hanping.

A formátum maga egy pillantásra áttekinthető (a színezés természetesen csak az érthetőséget segíti itt, hiszen nyers szövegről van szó):

```
舉辦 举办 [ju3 ban4] /rendez (eseményt, rendezvényt)/szervez/
```

Anélkül, hogy a fenti szintaxison változtatnék, a CHDICT-ben számos szemantikai kiegészítést teszek. Így például a zárójelezett szövegrészek metainformációnak számítanak, a keresésben nem vesznek részt, és bizonyos helyzetekben zárt címkelistáról kell származniuk.

A bejegyzések között, megjegyzésként jelölt sorokban kiegészítő információk állnak majd a soron következő szócikk státuszáról, korábbi verzióiról, a módosítások időpontjáról és szerzőjéről. Ez a kompatiblitiás megőrzésével eltérés az előképektől, mert a CHDICT adatfájlja így elsőként a teljes változástörténetet magában foglalja.

4.2 Licenc

A CHDICT anyagát Creative Commons licenc alatt teszem közzé. Ez részben a felhasznált anyagok licenceléséből eredő kényszer. Fontosabb azonban, hogy a közösségi licenc lehetővé teszi a tartalom továbbfejlődését abban az esetben is, ha az eredeti fenntartó magára hagyja a projektet. Nem utolsósorban pedig etikai szempont, hogy így a közreműködők azonos feltételek mellett megőrizhetik saját szerzői jogaikat az összes hozzájárulásukra, mivel a verziótörténet is az adat szerves részét képezi.

4.3 Közzététel

A munka elkészültével két végterméket teszek közzé. Az egyik a szótári tartalom, amely letölthető lesz mind a szótár weblapjáról, mind egy automatikusan frissülő

Github-repozitóriumból. A másik az említett weblap maga, amelynek forráskódja már most is elérhető egy Github-repozitóriumban.

A keresési funkciók túlmutatnak a kínai és magyar szavak megtalálásán, és a CHDICT szótári anyagát több más forrással ötvözik. Legfontosabb az automatikus kézírás-felismerés és az egy kattintással elérhető vonássorrend-animációk. Mindkét funkció Shaunak Kishore *Make Me a Hanzi* projektjén[10] alapul, amelyhez csekély mértékben magam is hozzájárultam. Apró részlet a kínai szófrekvenciák tekintetbe vétele a magyar keresési eredmények sorrendezésekor. Ugyanaz a célnyelvi szó gyakran több bejegyzésben is szerepel, amelyek közül célszerű a gyakoribb kínai szavakat előre sorolni, hogy a releváns találatok álljanak a lista elején.

Egyenrangú funkciója a weblapnak a nyilvános szerkesztői felület. Akárcsak a Wikipedia „laptörténet" fülén, a CHDICT weblapján is megtekinthető lesz a szótár összes változása, illetve egy-egy szócikk saját változástörténete.

A szerkesztőfelület is számos, nyelvi adatra épülő kényelmi szolgáltatást tartalmaz majd, így például az egyszerűsített címszó bevitele után automatikusan felkínálja az ismert hagyományos változatot és a pinyin-átiratot, illetve ha a címszó megtalálható a CEDICT-ben vagy a HanDeDict-ben, akkor az ezekben álló szócikket. Az efféle funkciók célja, hogy megkönnyítsék a szerkesztési munkát, ezáltal elősegítsék a szótár bővülését és fejlődését. A nyers szöveges adatformátum sem a keresés során, sem a szerkesztőfelületen nem jelenik meg eredeti formájában.

5 Összegzés és kitekintés

Az írásban bemutattam a CHDICT-en eddig végzett munkát, amelynek eredményeként haramosan egy 10 ezer szócikkes, nyílt, közösségileg szerkesztett kínai-magyar szótár indul útjára. Az intellektuális kihíváson túl leginkább izgalmas kísérletként tekintek a projektre, két kérdésre remélve választ.

Elsősorban: Lehetséges-e az elérhető nyílt forrásokra alapozva, belátható munkabefektetéssel létrehozni egy használható méretű és minőségű kétnyelvű szótárat? Ha igen, úgy a munka útmutatásul szolgálhat más nyelvpárok számára is.

Ami a minőséget illeti, a fenti kérdés megválaszolása kihívást jelent. A szótárfordítás nem bevett gyakorlat, s nem összevethető a gépi fordítás kiértékelésével, de az emberi fordítások minőségbiztosításával sem. Szúrópróbaszerűen kiválasztott szócikkek emberi értékelése, más szótárakkal való összevetése kínálkozik lehetőségként egy későbbi vizsgálat számára. Végeredményben azonban a választ a weblap látógatószáma adja majd meg. A HanDeDict weblapját 60-150 látogató keresi fel naponta, akik 500-2000 lekérdezést hajtanak végre. A CHDICT esetén a beszélők számából kiindulva ennek nagyjából a tizedére számítok.

Miután a szótár weblapja elindul, nagy értéket jelentenek majd a naplózott használati adatok. A gyakori lekérdezések kijelölik a bővítés irányát és a magas prioritással gondozandó szócikkeket is, illetve képet adnak arról, mennyire kielégítő a szótár aktuális terjedelme.

[10] https://skishore.github.io/makemeahanzi/

Jócskán van lehetőség a szótár proaktív fejlesztésére is. Kézenfekvő a magyar tulajdonnevek módszeres bevitele, amelyeket a kínai átiratok kiszámíthatatlansága miatt hasznos szerepeltetni. A fontos nyelvtani információk feltüntetése is további értéket jelent, ám ezeknél egyre kevesebb nyílt forrásra alapozhatunk.

A második kérdés, hogy a kezdeti állapot közzététele után életképes-e a közösségileg szerkesztett modell egy olyan „kicsi" és bizonyos tekintetben speciális nyelvpár esetén, mint a kínai-magyar. Ha igen, úgy átültethető-e vajon a modell más, vélhetően nagyobb impaktfaktorú, eltérő kihívásokat és elvárásokat támasztó nyelvpárokra, mint például az angol-magyar? Erre a válasz megjósolhatatlan.

6 Eddig közzétett anyagok

A CHDICT teaser-oldala: https://chdict.zydeo.net
Az összes forrást ötvöző XML-fájl a kezdeti szótárfordításhoz:
https://chdict.zydeo.net/files/backbone.zip
A webes alkalmazás forráskódja: https://github.com/gugray/ZydeoWeb
Az élő webes alkalmazás, amelyen a HanDeDict kínai-német szótár kereshető (illetve hamarosan szerkeszthető): https://handedict.zydeo.net/
A szótár fordításához használt egyedi alkalmazást szívesen az érdeklődők rendelkezésére bocsájtom.

Hivatkozások

1. Bartos, H., Imre, H.: Kínai-magyar szótár. Balassi Kiadó (1998)
2. Breen, J.W.: Building an Electronic Japanese-English Dictionary. JSAA Conference, Brisbane (1995)
3. Cai, Q., Brysbaert, M.: SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. PLoS ONE 5(6): e10729. doi:10.1371/journal.pone.0010729 (2010)
4. DeFrancis, J., Zhang, Y., Mair, V. (eds.): ABC Chinese-English Comprehensive Dictionary. University of Hawai'i Press (2003)
5. Naszódi, M.: Statisztika megbízhatóság a nyelvészetben. Széljegyzetek egy szótárbővítés ürügyén. In: Tanács, A., Varga, V., Vincze, V. (eds.) XI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 34-45. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2015)
6. Tschirner, E.: Häufigkeitsverteilungen im Deutschen und ihr Einfluss auf den Erwerb des Deutschen als Fremdsprache. In: Marello, Carla a.o. (eds): Atti del XII Congresso Internationale di Lessicografia. Alessandria (2006) 1277-1288.
7. Apel, U.: Ein elektronisches japanisch-deutsches Wörterbuch auf Datenbankbasis – Über das Finden von Wörterbucheinträgen im Computer-Zeitalter. In: Gössmann, Hilaria; Mrugalla, Andreas (eds): 11. Deutschsprachiger Japanologentag in Trier (1999) Bd. II, pp. 627–644.