Statisztika megbízhatósága a nyelvészetben
Széljegyzetek egy szótárbővítés ürügyén

Naszódi Mátyás

MorphoLogic, 1122 Ráth György utca 36.
naszodim@morphologic.hu

Manapság szinte korlátlan mennyiségben lehet természetes nyelvű szövegeket elérni a www jóvoltából. Emiatt a nyelvi kutatásoknál, eszközök fejlesztésénél erősen támaszkodnak nyelvi statisztikákra. A megbízhatóság kérdésével viszont kevesen foglalkoznak, pedig ez kulcskérdése a tömeges adatok felhasználhatóságának. Ez a cikk azzal foglalkozik, milyen jellegű objektív korlátai vannak a statisztikáknak, és hogyan lehet becsülni a megbízhatóságot.

1 Bizonytalanságok a nyelvben

A 80–as években, mikor természetes nyelvű szövegek feldolgozásával kezdtem foglalkozni, matematikusként azt a tényt kellett tudomásul vennem, hogy semmi sem százszázalékos. Korábban olyan témakörben dolgoztam, ahol egy állítás vagy igaz, vagy nem, esetleg az adott axiómarendszerben nem eldönthető. Nem így a nyelvészetben. A természetes nyelv tele van gyengén definiált fogalommal, szabályokkal, melyekre mindig találunk kivételeket, többértelműséget, redundanciát**.**

1.1 Bizonytalanságok szószinten

Már az a kérdés, hogy egy szó magyarnak tekinthető, vagy esetleg a karaktersorozat hibás, nem egyszerű kérdés. Ha nem lettek volna a tizenkilencedik századi nyelvújítók, nem lenne egységes írásunk, *mozdony* szavunk. Ha Karinthy nem írta volna le a *patyolat* szót, azt sem ismernénk. Ha Kellér Dezső nem alkotta volna meg a *maszek* szót, nem használnánk. Egy újságcikkben megjelenhetnek tulajdonnevek időnként nem magyar abc betűit használva. Idegen szavak kerülnek a köznyelvbe, melyeknek lenne ugyan magyar megfelelője, de mégis a jövevény terjed el. Ezeregy oka lehet, hogy egy-egy új szót bevesz a nyelv, másokat elfelejt. A toldalékolás is változik, illetve bizonytalan. A régies múlt *kihala vala* a köznyelvből. Megállapodott szabályokat rúg fel a gyakorlat, illetve fel nem ismerhető régi szabályok felülírnak szokásosakat. (*gyorsan*, *boldogan*, de *nagyon*, *fiatalon*, *gazdagon*; *mondta*, vagy *mondotta*; *falsság*, *nyersség*, de *rosszaság*, *gyorsaság*, sőt *frissesség*, *bölcsesség*; *gondtalan, de gondatlan*) De nem csak időbeni változások léteznek. Ami megengedett egy szaknyelvi szövegben, helytelen lehet egy köznapi mondatban.

A szavak jelentése sem egyértelmű. Egy szóalaknak nem csak azért lehet több értelme, mert azonos alakú, lényegesen más szóról van szó (*lépnek* FN, *lépnek* IGE; *értem = érik* IGE+ME1 */én+ért/ ért* IGE+E1 tárgyas...), hanem egyes képzett, összetett szó új jelentést kaphat, de megtarthatja eredeti nyelvtani struktúrából eredő jelentését is. (*lovagol – lovon közlekedik / lovagol valamin – ragaszkodik egy érvhez*). Számos más jelenség is okozhat többértelműséget. Az esetek többségében az egyik értelmezésnek sokkal nagyobb a valószínűsége, mint a többinek, ha másként nem, a szövegkörnyezet függvényében. Angol nyelvben a szavak többértelműségének többsége abból ered, hogy a szavaknak névszói és igei jelentése is van, de ezt a szórend egyértelműsíti. Az erősen ragozó nyelveknél a szóalakok nagyszámú változata miatt keletkeznek különböző módon generálható, de azonos alakú szavak.

1.2 Bizonytalanság a nyelvtanban.

A mai nyelvgyakorlat elüt a korábbitól. Mostanában számos angol nyelvből átvett forma jelentkezik a hétköznap és a sajtó gyakorlatában. Ezen túl nyelvészeti cikkeket olvasva sok olyan mintapéldát találtam, melyet a szerző helytelennek, esetleg kérdésesnek talál, nekem meg nyelvileg tökéletesnek mutatkozik, és viszont, helyesnek jelzett mondatban találok javítandó hibákat. Néha egymásnak ellentmondó szabályokat kell egy mondatra alkalmazni. A mondatok nyelvtani elemzése köztudottan nem egyértelmű. A gépi elemzők nagyságrenddel több alternatívát tárnak fel, mint az ember feltételezi olvasás közben. Ezek nagy hányada fel sem merül egy olvasónak, mert a szemantikai megkötések, az ismert és gyakori minták elnyomják a ritka lehetséges elemzést. Az itt is igaz, hogy az esetek többségében egy-két elemzés dominál, a többinek kis súlya van. Arról már nem is beszélek, hogy egy nyelvre több lehetséges nyelvtant lehet készíteni. Míg a szavakról, minősítésükről egyöntetűbb rendszerek vannak, a nyelvtan, főként a formalizált nyelvtan, szerzőnként eltér. A nyelv egy objektív jelenség, de a nyelvtan ember által alkotott modell, ami lehet jó, pontos, de sohasem a valóság maga[5]. Azért alkotja a nyelvész, hogy áttekinthetőbbek, kezelhetőbbek legyen a nyelv jelenségei.

1.2 Bizonytalanság jelentésben, fordításoknál

Egy mondatnak számos interpretációja lehet egy másik nyelven, illetve szemantikai reprezentációban. Léteznek jó és gyenge fordítások. Ritkán beszélhetünk tökéletesről, de megfelelő fordításról gyakran. Ebben az esetben olyan valószínűségi modellt lehet alkalmazni, ami szerint azt mérjük, hogy két különböző nyelven írt mondatnak mi a korrelációja, vagyis ugyanabban a helyzetben, ahol az egyik elhangzik, a másik nyelven milyen valószínűséggel hangzik el a másik. Mivel a mondatok száma gyakorlatilag végtelen, ezt nem lehet számba venni, viszont az egyes mondatoknál kevés a minimális valószínűségi küszöböt elérő mondatpárok száma. Ezt használják ki a statisztikai és memória alapú fordítók.

Sorolhatnék még számos bizonytalansági kérdést a természetes nyelveknél. Tulajdonképpen a természetes nyelv velejárója a nem teljesen meghatározottság, illetve a többértelműség [5]. A nyelvészeti kérdésekre általában nem tudunk igennel, nemmel válaszolni. Megfelelőbb az olyan valószínűségi modell, melyben a legtöbb esethez nagy (egyhez közeli) vagy kicsi (nulla körüli) valószínűséget rendelünk. Számos esetben nem tudjuk minősíteni határozottan a nyelvi jelenséget. Ezért érdemes valószínűségi modelleket alkalmazni.

2 A nyelvi statisztikák karaktere

Ezek dacára nem ismerek olyan helyesírás-ellenőrzőt, amely nem kategorikus választ ad arra, hogy egy szó helyes vagy nem. A stílusellenőrzők zöme is határozott állítással ítél, esetleg egy-két esetben ad figyelmeztető jellegű az üzenete. A fordítóprogramok sem zavarják a felhasználót azzal a közléssel, mennyire biztos a szöveg interpretációja, esetleg hány száz, ezer egyéb alternatívát ismer az adott mondat áttételére.

A tapasztalat azt mutatja, hogy nyelvi esetek túlnyomó többségénél a bizonytalanság karakterisztikája olyan, hogy a gyakori esetek elnyomják a ritkábbakat. Értsd ezen azt, hogyha két-három választék van a megoldásra, az egyik általában nagyon gyakori, a többi ritka. Úgy is lehet magyarázni, hogy ha egy konfidenciatartományt jelölünk ki, akkor ebbe kevesen jutnak be. Ha sok lehetséges eset van, akkor azok kis százaléka lefedi a futó szövegek nagy százalékát. Ez egy közismert jelenség, amely igaz természeti törvényszerűségekből eredhet.

2.1 A Zipf-törvény

Számos természetben előforduló sokaságnál igaz a következő összefüggés: Ha megkülönböztethető csoportokat alkotunk az előforduló egyedekből, és a csoportokat előfordulásuk gyakoriságának sorrendjében rendezzük, akkor az n–edik csoport előfordulásának relatív gyakorisága a következőképp becsülhető:

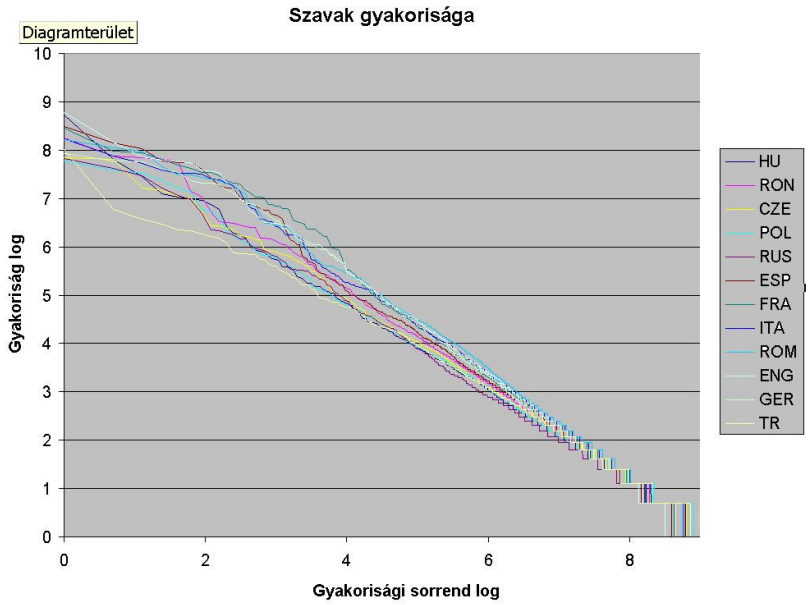
$$f(n) \approx C/n^s \qquad (1)$$

ahol C egy normáló konstans, s pedig egy egynél kicsit nagyobb kitevő [1].

A törvény nagyobb sokaságoknál igaz. Ha azt nézzük, hogy a vagyon hogy oszlik el az emberek közt, vagyis gazdagság szerint sorrendezzük az embereket; a népesség hogyan oszlik el a lakóhelyeken, vagyis a településeket sorrendezzük a lakosok száma szerint, hasonló jelenséggel találkozunk. A lényeg, hogy a csoportok száma meghaladja az ezret.

Kis n–nekre, tehát a leggyakoribb csoportokra nem jó ez a becslés. A leggazdagabbak eloszlása sem igazán követi a képletet, ott más törvények is belejátszanak az adatok alakulásába. Ha például a szövegekben előforduló szavakat vagy szóalakokat vizsgáljuk, akkor a leggyakoribb szavak relatív gyakoriságára nem állja meg a helyét a becslés, a ritkább értékeknél sem használható a képlet – a mérési, minta-

vételezési hiba miatt – de a legalább tízszer előforduló egyéb szavaknál elég jó a közelítés.



1. ábra A szavak gyakorisági grafikonja jól mutatja, hogy a középmezőnyben követi a Zipf törvényt.

Az s kitevő értéke függ a sokaság típusától. Ha a szógyakoriságot nézzük, akkor magyar nyelvnél ez közelebb van az egyhez, mint az angolnál. Ha megnézzük a függvény grafikonját, monoton csökkenő, alulról konvex függvényt kapunk, ami az ordinátához simul. A nyelvi esetekben az s annál közelebb van az 1–hez, minél nagyobb az elvi csoportok száma. Ez magyarázza a magyar és angol közti eltérést. 1 nem lehet az s, mert akkor nem lehetne normálni, mert a görbe alatti terület korlátlan.

1. Táblázat: s becsült értéke Orwel 1984 fordításai alapján

Nyelv	Szavak	Szóalakok	>1	>2	>3	s
magyar	88054	19236	6575	3378	2638	0,9327
török	73224	19186	7046	4164	2944	1,0059
orosz	74109	18560	6443	3873	2485	0,9284
cseh	79681	17630	6673	3991	2817	0,9832
lengyel	80186	19598	7399	4237	2898	0,9411
roszin	89560	16874	6651	3963	2863	1,0059
francia	111395	11367	5629	3748	2842	1,1586
spanyol	95380	11442	5390	3513	2602	1,1384
olasz	102578	13771	6321	4025	2983	1,1060

román	107477	14154	6479	4217	3100	1,0754
angol	104759	9211	4906	3397	2639	1,2258
német	100052	14048	5754	3598	2628	1,2502

Az s–et tapasztalati mérések alapján becsülik, de ennek a becslésnek is vannak megbízhatatlansági tényezői. Ha viszont az s-et ismerjük, becsülhetjük a teljes sokaság számát is. Például a szavak, szóalakok számát egy nyelven. A becslés nem pontos, de nagyságrendi adatokat meg lehet állapítani. A fenti táblázat adatai nem a nyelvről, hanem a regényről, illetve a fordításról/fordítóról adnak tájékoztató adatot.

A becslés nem alkalmazható egyszerű betűstatisztikára betűírás esetén, mert az egy nyelven használt elemek száma erősen ezer alatt van. Ideografikus (pl. kínai) írás esetén viszont már megfelelő az alaphalmaz számossága. Általában digráf statisztikára sem igazán jó, de a lehetséges betűhármasok száma egy nyelven már kellően nagy, hogy megkísérelhessük a Zipf-képlet konstansainak meghatározását.

2.2 A nyelvi statisztikák abszolút pontatlansága

Minden nyelvre jellemző a betűstatisztikája. Már itt is gondban lehetünk, mert ha például megnézzük Arany János és Petőfi Sándor írásait, akkor a két azonos korban élő költőnél eltérések fedezhetők fel. Arany János nagy hangsúlyt helyezett arra, hogy mély hangrendű szavakat használjon, emiatt az *a* és az *e* betűk gyakorisága alig különbözik, míg Petőfinél lényegesen több az *e* betű. Tehát egy nyelvnek nincs jól meghatározható betűvalószínűsége, illetve ha azt feltételezzük, hogy van, akkor nehéz az általános statisztikának megfelelő mintavételezést alkalmazni.

Ha meg akarják határozni, hogy hány és milyen hal van a Balatonban, akkor azt úgy teszik, hogy egy mérhető területen lehalásszák a halakat, és az itt nyert eredményből extrapolálnak. Nem mindegy viszont, hogy ezt hol teszik. A part közelében más az eredmény, mint a parttól távol, és Siófoknál eltér az összetétel a keszthelyi öblétől. Nevezzük ezt a bizonytalanságot ***haleffektus****nak.*

Habár Arany Jánosnál és Petőfi Sándornál eltérnek a statisztikák, mégis nagy vonalakban hasonló az eredmény, A hasonlóság azt jelenti, hogy a gyakoribb betűk relatív gyakorisága legfeljebb 20-30 %-kal tér el egymástól a két írónál és általában a magyar szövegekben. A ritka betűknél ennél nagyobb eltérés is lehetséges.

Az egyszerű karakterstatisztikáknál sokkal jobban jellemzi a nyelveket, milyen karakterek követik egymást, ezért betűkettősök, betűhármasok felmérésével elég jól meg lehet határozni a nyelvet, melyen a szöveg íródott [2]. Itt sem abszolút kiértékelésről van szó. A módszer inkább az, hogy a kérdéses szöveg gyakoriságvektorát veti össze a szóba jövő nyelvek előzetesen elkészített gyakoriságvektorával, és amelyikhez a legközelebb áll, azt a nyelvet kiálltja ki győztesnek. A módszer tehát a Bayes-módszer egyik alkalmazása.

Ha nem karakter-, hanem például szóstatisztikát készítenek, akkor az eltérések jelentősebbek. A leggyakoribb szavak nyelvtani funkciós szavak. Például a névelők, kötőszavak. Ezek aránya minden szövegben hasonló. A továbbiak viszont eltérnek. A sorrendben első húsz-harminc szó minden nagyobb szövegben előfordul, ráadásul

egyenletesen eloszolva a szövegtörzsben. A gyakoriak, de nem az élbolyban szereplők már nem. Egy vaskosabb receptkönyv elemzésénél vettem észre, hogy a könyv első felében a *só* majdnem annyiszor szerepel, mint a másodikban, de a *cukor* a második felében sokkal többször. Hát persze, mert az édességek a könyv végén szerepeltek.

Ez azt jelenti, hogy a szavak gyakoriságát a **haleffektus** miatt nehéz becsülni. Inkább csak kvalitatív, mintsem kvantitatív eredményeket várhatunk. Ilyen eszközökkel viszont már szerzőket is fel lehet ismerni, illetve szaktémákat lehet megkülönböztetni. Ennek finomított változata elég a plágiumkereséshez.

Érdekes, hogy az azonos jelentésű szavaknál hogyan lehet megkülönböztetni az alaktól a szót. Füredi Mihály Magyar nyelv szépprózai gyakorisági szótárában[1] az *az* névelő és az *az* mutató névmást külön számlálta. Miután ilyen munkát csak egyszerű gépi módszerrel lehet elvégezni – a nyolcvanas években gépi egyértelműsítőről legfeljebb csak álmodhattunk – egyszerű emberi beavatkozással oldották meg a problémát: mivel mindkettő értelmezés gyakori a magyar nyelvben, ezért elég egy részmintában emberi erővel elvégezni a statisztikát, és a két értelmezés arányát fixnek véve az alakok alapján lehet jól becsülni a kettő gyakoriságát a teljes korpuszon. De mi van, ha az egyik értelmezés ritka, mint a *meg* igekötő (gyakori) és kötőszó (ritka) használatánál. Minden esetet megnézni nem lehet emberi erővel, tehát nem tudhatjuk, mi a ritka változat gyakorisága. Ha mindkettő ritka, akkor persze átnézhetőek az adott szóalak előfordulásai, emberi döntéssel megadható, hol, melyik a helyes értelmezés. Ilyen szóból viszont – a Zipf-törvénynek megfelelően – rengeteg van. Tehát az eseti döntés megoldható, de a tömegfeldolgozásuk nem.

A tapasztalatom az, hogy a nyelvi adatok abszolút pontossága egy anyagon belül nem függ az eset gyakoriságától, tehát állandó. Akár gyakori esetről van szó, akár ritkáról, a tévedés gyakorisága ugyanaz, tehát a relatív tévedés fordított arányba van az illető gyakoriságával. Emiatt gyakori esetekről sokkal megbízhatóbbak az adataink, mint a ritkákról.

2.3 Anyaggyűjtés matematikája

A nyelvi feladatok többsége olyan, mint egy szótár készítése. Szavakat gyűjtünk, és hozzárendelünk információkat. Amennyire lehet, ezt géppel vagy gép segítségével végezzük. A szótár – ha helyesírás-ellenőrzőről van szó – csak szógyűjtés. Erősen ragozó nyelveknél további osztályozás is nélkülözhetetlen. A szógyűjtés manapság abból áll, hogy nagy korpuszokban fel nem dolgozott szavakat, kifejezéseket keresünk, majd feldolgozzuk azokat. A gyakori szavakra gyorsan rátalálunk, és feldolgozása is egyszerűbb, hisz az ember a felmerülő kérdésekre biztos választ tud adni. A ritkábbakra nehezebb rálelni, és nem mindig egyszerű a szóhoz rendelt tulajdonságokat pontosan megadni.

Tehát a szótár bővítése annál nehezebb, minél ritkább szavakkal kell törődni. Ha ezt nem vennénk figyelembe, akkor a szótárbővítés sebessége a tételszámokkal lineárisan nőne:

$$T(N) = \sum_i e_i = N*e \qquad (2)$$

ahol N a feldolgozott szavak száma, e pedig az egy szóra fordított idő. Ha így lenne, akkor is gond a szövegek lefedettsége, mivel a Zipf-törvény alapján a szöveg nagy hányadát ugyan lefedhetjük a gyakori szavakkal, de ha még nagyobb lefedettséget akarunk elérni, akkor hatványozottan nagyobb korpuszokat kell vizsgálni.

Emiatt a szótár mennyiségével rohamosan nő az elvégzendő munka. Új elem megtalálásához a gyakoriság reciprok arányában nő a szükséges korpusz mérete, így a feldolgozás ideje a tételszámokkal legalább négyzetesen nő:

$$T(N) = \sum_i e_i = C*\sum_i i^s = C* N^{1+s}/(1+s) > O(N^2) \quad (3)$$

Ráadásul ahhoz, hogy javuljon a lefedettség – a ritkább elemek kevéssel javítják ezt a paramétert – a pontosság növeléséhez ezen érték hatványát kell munkába fektetnünk. Persze a talált objektumról véleményt is kell mondani. Ha csak helyesírás-ellenőrzőről van szó, akkor meg kell ítélni, az újonnan talált szó eleme-e a nyelv szókészletének, vagy sem. Ez azért fontos, mert nem minden meglelt sztring helyes szó. Még az sem segít, ha a gyakoriságát nézzük. A *hüje* szó ötször több esetben szerepel a google szerint a szövegekben, mint például az *oktondi*. Ennek ellenére az előző antiszó, míg az utóbbi teljesen köznyelvi, helyes alak. Ha feltételezzük, hogy egy szó felvétele, osztályozása is fordított arányban van a gyakoriságával, akkor ez a korábbi képletet erősíti. Így egy minőség előállításához szükséges idő nagyságrendjének becslése:

$$T(q) > O(1/q^2) \quad (4)$$

ahol $1-q$ a lefedettség, tehát valamilyen minősítéshez tartozó mérőszám.

2.4 A tévedések matematikája – a mennyiségi korlát

A Zipf-törvény magyarázataiban szerepel az is, hogy a kis valószínűségű osztályok értékei nem igazán követik a képletet. Ez nagyszámú osztályok esetén természetes, hisz a ritka elemek gyakorisága 1 körüli, vagy annál kisebb, akkor – egész értékű előfordulás miatt – eleve létezik egy pontatlanság. A pontatlanságnak más okai is vannak:

1. a korábban említett mintavételezési hiba – a haleffektus
2. az emberi döntések hibája
3. a forrásanyagban (korpuszban) fellelhető hibák

Ha ezt is beszámítjuk, akkor a szótár minősége – az egyedek helyes és helytelen felvételének aránya – fokozatosan romlik a szótár növekedtével. Ha a szótárkészítő egyenletes minőségben dolgozna, akkor a szótár minőségének felső határa a szótárkészítő helyes döntéseinek arányával azonos. Ha viszont feltételezzük, hogy a helyes döntés aránya csökken a felvett szó gyakoriságának csökkenésével, akkor azt a képletet kapjuk, hogy van egy olyan határ, amikor a bővítés már ront. Ha valaki nem hiszi, akkor nézze meg egy keresőben, hányszor szerepel a *kéttannyelvű* szó a weben, és hányszor a *két tannyelvű* helyes alak. A hibás sokkal többször. Ez a konkrét példa nem szerepel ugyan a helyesírási szótárakban, de az általános nyelvtani szabályok miatt külön kell írni. Nem úgy, mint a *négykerék-meghajtású*t. (Házi feladat, miért?)

És e két példa nem az igazán bonyolult eset a magyar nyelvben. Vagyis bárhogy állapítjuk meg a szótár minőségi követelményét, egy mennyiség elérése után már rosszabb eredményt kapunk a követelménynél.

A minőséget persze lehet javítani. Az egyik lehetséges módszer, hogy a döntéseket két, esetleg több független (kis korrelációval rendelkező) döntnök (algoritmus) hozza. Ha kicsi a hibaszázalék, ez két független döntés esetén közel felezi a hibás kódolásokat. Mivel a hibák aránya a 3. képlet alapján négyzetesen nő a gyűjtés mennyiségével, előbb utóbb akkor is objektív korlátba ütközik a minőség megtartása, és ezen a sok független döntnök sem segít.

Ha például szótárkészítésről van szó, akkor tapasztalatom szerint egy ember belátható idő alatt maximum 10 000 tételből álló szótár készítésére képes. A nagyobb szótárakhoz közösségekre, lektori rendszerekre van szükség. A nagy – 100 000-200 000 vagy több tételt tartalmazó szótárak többgenerációs munkák. Tehát a minőség eléréséhez sok ember paralel döntésére van szükség. Ha egy ilyen szótárat módosítani akarnak, akkor a hagyományos szócikkiosztás módszere nem célravezető, hisz ilyenkor már ritkább szavak kerülnek sorra, és lehetséges, hogy többet rontunk a meglévő készleten, mint javítunk. Ez látszik is az utóbbi időkben megjelent összevethető szótárak minőségén.

Érdekes probléma például a helyesírás-ellenőrzők adatbázisának építése. Nem feltétlen a nagyobb szótár a jobb. Lefedettségben ugyan igaz, de a helyesírás-ellenőrzőnek az a feladata, hogy hiba esetén jelezzen. A hiba pedig azt jelenti, hogy a szó a szövegben helytelen. Ha minden helyes formát megengednénk, akkor olyan elütések, melyek helyes szóalakhoz vezetnek, elrejtik a hibát. Ilyenből rengeteg van. Például a *tanit* helyes nyelvileg, *tan+i+t*, de nem gyakori az előfordulása. Ezzel szemben a *tanít* elég gyakori, és a hibás, rövid *i*-vel történő írása is gyakrabban fordul elő, mint a korábbi eset.

Egy ellenőrzőprogram jóságát abban szokták mérni, mennyi az elsődleges és másodlagos hibák száma. Elsődleges hiba, ha nem ismer fel egy helyes szót, másodlagos, ha jónak tekint egy hibásat. A két hibának más a súlya. Tehát a minősítés (az eszköz rosszaságának mérőszáma) lehet a következő:

$$M = C_1*E_1+C_2*E_2 \tag{5}$$

ahol a C-k a súlyok, és az E-k a hibák relatív gyakoriságai. Általában a másodlagos hibát nagyobb súllyal szokták számítani, mint az elsődlegest. $C_1<<C_2$. A felmérést a legritkább esetben végzik futó szövegeken, inkább csak kigyűjtött szókészleteken. A szóellenőrzők nem képesek figyelembe venni a szövegkörnyezetet.

Ennél finomabb súlyozás logikusabb lehet. Tudniillik a konkrét hibáknak más a hatásuk a szövegekben. Ha azt írjuk, hogy *hüje*, vagy *talpalattnyi*, az nem okoz akkora galibát, mintha a *kóros* helyett *koros* kerül a szövegbe. Míg a korábbiaknál az olvasó helyesen értelmezi a hibás karakterláncot, az utóbbinál értelemzavaró lehet az elírás. Ezért ha egy helyesírás-ellenőrzőbe felveszünk egy új szót, akkor azt is meg kell vizsgálni, hogy milyen gyakori szavakhoz van közel a szó vagy a szóból előállított szóalak, és ez mekkora gondot jelenthet. Ezért nem szerepel egy helyesírás-ellenőrzőben sem a *suly* (*betegség*) szavunk. Egy új szó felvételének haszna a következő:

$$M = C_1*E - \sum_j C_{2j}*m_j - \sum_i C_{3i}*m_i \quad (6)$$

ahol C_{2j} a másodlagos hiba kára, ha a szó helyett a j-edik szó kerül a szövegbe, C_{3i} a másodlagos hiba kára, ha a j-edik szó helyett az új szó kerül a szövegbe, a két szó közti távolság pedig m_{ij}, vagyis arányos annak a valószínűségével, hogy az i-edik szó helyett a j-edik írjuk le.

A szótár minőségre a következő becslést adhatjuk.

$$M = \sum_i C_{1i}*E_i - \sum_i \sum_j C_{2ji}*m_{ij} \quad (7)$$

ahol C_{2ji} a másodlagos hiba kára, ha az i-edik szó helyett a j-edik szó kerül a szövegbe, a két szó közti távolság pedig m_{ij}.

Összefoglalva, a több gyakran vezet minőségromláshoz, túl a kódolási hibákon. Az a mennyiség, amelyen felül már nem javul a nyelvi gyűjtemény, függ a gyűjtemény típusától, a kódolás módjától, de munka közben becsülhető. Ezen túl nem szabad menni, mert többet árthatunk, mint használunk!

2.5 A fordítások minősítése

Fordítások esetén – ha fordítómemóriáról vagy statisztikai fordítóról van szó – két nagy eltérés van az emberi, fordításhoz képest:

1. Az adatokat nem ember gyűjti, hanem meglevő korpuszok szolgáltatják, melyet általában kontroll nélkül fogad be a rendszer.
2. Az eredményekről nem igen-nem jelleggel minősítünk, hanem valamilyen jósági mértéket lehet, kell mondani.

Mindkét esetben a gyakori mondatok – amelyek fordítása már bekerült a rendszerbe – jól interpretálódnak abban az értelemben, hogy nagy valószínűséggel a módszer kiválasztja a megfelelő interpretációt. A gond ott van, hogy a lehetséges mondatok száma sok nagyságrenddel meghaladja a nyelvben található szavak számát. Ha a Zipf törvényét nézzük, és szópárok, esetleg szóhármasok eloszlását vizsgáljuk egy nyelvben, akkor várhatóan az s konstans nagyon közel van az 1-hez, ennek következtében arra a kérdésre, hogy a gyakoribb osztályok mennyire fedik le a futó szövegben előforduló esetek felét, kétségbeejtő választ kapunk. A ritka elemek lesznek nagyobb hányadban. Ha ez igaz, márpedig így van, akkor a statisztikáink hibája nagyon nagy, akár a haleffektusról, akár a korpuszban fellelhető hibáról van szó. A fordítás ráadásul sohasem lehet tökéletes.

A fordítás minősítését is szeretnénk gépesíteni. Erre az általánosan használt módszer a statisztikai kiértékelés, pl. a BLEU [4]. A gond ezzel ott van, hogy a kiértékelés pont azokat a paramétereket nézi, amilyen alapon generálják a fordított szöveget, tehát részrehajló. Ha csak emberi fordításokat értékelnének ki ezzel a módszerrel, akkor persze objektívabb lenne, hisz a kiértékelés módja korrelálatlan a fordítás módszerétől. [5] Ezt igyekszik kikerülni az ITRANSLATE4 projekt. A módszer egyszerű. A felhasználók visszajelzése értékeli a fordítást. Ennél jobbat csak szakemberek bevonásával lehet elvégezni, ami költséges, és nagy mennyiség kiértékeléséhez kivihetetlen.

Bár a statisztikai fordítás hívei elméletileg támasztják alá, hogy a feldolgozott korpusz méretével egyre jobbak a fordítások, és elvileg – korlátlan forrást feltételezve – a minőség is tökéletesedik, a gyakorlat, és az általam vázolt képletek alapján ez nem igazolódik. Nem beszélve arról, hogy a felhasználható korpuszok mérete sem korlátlan, emiatt a (4)-es képletnél is rosszabb a helyzet. A különböző nyelvek közti fordítás minősége különbözik. A különbség alapvetően nem attól függ, melyik fordítóprogramot használjuk – a google-ét a bing-et, vagy bármi mást, hanem a nyelvpártól. Persze az is számít, mekkora anyag van a fordító tarsolyában. A statisztika általában annál pontosabb, minél több anyag van mögötte, de a következők sokkal inkább meghatározók:

1. Milyen nagy a szóformák készlete az adott nyelven
2. Milyen közel van a két nyelv nyelvtana
3. Vannak-e javító trükkök

Ebből a szempontból a magyar nehéz helyzetben van. A fordításnál a másik nyelvtől nagyon különbözik. A szóformák száma nagyságrenddel nagyobb, mint más nyelveknél. Meg is látszik az eredményen. Míg egy angol-francia fordításnál a statisztikai fordítók minőségével a hétköznapi felhasználók elégedettek, a magyar-angol esetben gyengébb minőséget kapunk.

2 Javítási lehetőségek

Ha ennek ellenére jobb minőséget akarunk elérni, akkor a következőket lehet tenni. A sokaság méretével csökken a Zipf-képletben szereplő `s` kitevő, vagyis a ritka elemek viszonylagosan kevesebben lesznek. Ha így csökkentjük az osztályok számát, akkor jobb eredményt tudunk elérni.

a. Ilyen lehet, ha szakszövegek, terminológiai kötések miatt az általános értelemben ritkább szavak egy részének gyakorisága megnő, de a ritka köznapi szavak nem kerülnek a feldolgozandók közé. Ekkor a szakszavakra összpontosítva bővíthetjük a szótárat. Persze ezzel csökkentjük a felhasználható korpuszok számát is.

b. Hasonló az eset az általános szóellenőrzőknél a sajátszótár használata esetén. Egy jobb tollú felhasználó is csak a tizedét, századát használja a nyelvnek, ezért, ha felmerül egy újabb helyes, de korábban nem szereplő szó, felveheti a szótárába. Egy ember nyelve mindig kisebb a teljes köznyelvnél, ezért az ő nyelvének lefedettségét így lehet biztonságosan bővíteni.

c. A módszerben az osztályok számát csökkentjük. Például egy helyesírás-ellenőrzőnél nem szóalakokat veszünk fel, hanem alapszavakat, melyekből szabályos módon generáljuk a szóalakokat. A ragozó nyelvek esetén e nélkül nem is lehetne jó helyesírás-ellenőrzőt készíteni, mert a szabályos szóalakok száma olyan sok, hogy a korábban említett kritikus mennyiség esetén a hibaszázalék meghaladja a helyesírás-ellenőrzőknél elvárt 5%-ot. Míg angol, francia, spanyol nyelveknél 100 000-200 000 szóalak felvételével a 97%-os lefedettség könnyen biztosítható, magyarban 1000 000 000 szóalak esetén sem érhető el 90%-nál jobb lefedettség.

3 Mérési tapasztalatok

Megkíséreltem az elméletet összevetni a valósággal. Különböző nyelven írt azonos – a mai köznapi nyelvhez közel álló szöveget kerestem. Erre alkalmasnak tűnt Orwell 1984 című műve. Az *újbeszél* szavain kívül a nyelvezete megfelel. Többek között a MULTEXT [7] projekt korpuszának is fő darabja volt. A különböző fordítások elérhetőek voltak. Többségük PDF formátumban, mások MS Word dokumentumként.

Az anyagot saját eszközökkel csupaszítottam le tiszta szövegállományra. Kihagytam az oldalszámozást, fejezetjelöléseket, lábjegyzeteket, elő és utószavakat, mert azok tartalma eltért a kiadványokban. Arra voltam kíváncsi, milyen mérhető eltéréseket tapasztalok nyelvenként.

Az eredmény kicsit váratlan volt. Azt vártam, hogy egyszerű becslésnél már határozottan kimutatható a Zipf konstansaiban az eltérés. A különbség meg is mutatkozott [1. ábra], de nem volt annyira karakterisztikus. Az `s` értéke – bár 1 körüli volt, de általában kisebb a vártnál. [1. táblázat]. Utólag magyarázni is tudom.

1. A konstansok általam számított becslése bár megbízható, de nem torzításmentes. Sokkal nagyobb minta kell, hogy a becsült érték közelebb legyen a valósághoz, vagy meg kell találni a torzítatlan becslés képletét.
2. Az értékek erősen függtek a fordítás minőségétől, tehát nem csak a regényt jellemezték, hanem a fordító választékosságát is.
3. Az értékek függtek a szöveg kódolásának minőségétől. A hibás, trehány kódolás miatt több ritka szóalak szerepelt, mint amennyi valójában kellett volna lennie. Ez különösen a roszin nyelvű anyagon látszott.

Azt gondolná az ember, hogy a fordítás miatt határozott szinkronszövegeket talál. Ez bizonyos értelemben így is van, de nem teljesen. Ha például csak azt számoljuk, hogy a főhős, Winston neve hányszor szerepel a műben, akkor mindegyikben 500 körüli, de 10%-os eltérés is tapasztalható. A legkevesebb fajta szóalakot persze az angolban és a németben találtam, és az erősen ragozó nyelvekben többet. Korábban abban a hitben voltam, hogy a finnugor nyelveknél markánsabb lesz a fölény, de nem így volt. A szláv nyelveknél talán a cseh a vezető. Archaikusabb, mint az orosz. Ez nem abban nyilvánul meg, hogy nem hat névszói esete van, hanem hét, hanem abban, hogy a birtokviszonyt is melléknévi ragozott alakban fejezi ki, akár az ukrán.

A grafikon ritka szavaira vonatkozó görbeelnyúlásra jobban lehet következtetni azok arányából, mint a teljes görbéből. Itt határozottabb különbség mérhető a nyelvek között.

4 Összefoglalás

Az elmélet és a tapasztalat is alátámasztja, hogy a nyelvészeti eszközök minőségének objektív korlátai vannak, melyet lehet becsülni még az eszköz készítése előtt, esetleg közben. A becslés nem mindig egyszerű. Összetettebb esetekben a munka során ellenőrző méréseket kell végezni, érdemes-e folytatni a munkát, vagy le kell-e állni,

esetleg módszert kell váltani. A minőségi korlát statisztikai alapon becsülhető, és ha nem vesszük figyelembe, a több munkával ronthatunk a készülő eszköz minőségén.

A méréseket érdemes nagyobb korpuszokra is kipróbálni, valamint nyelvi eszközökön, adatbázisokon megnézni, mik a számított korlátok.

A jövőben megkísérlem mérni, igaz-e a feltételezésem: a nagy memóriafordítók elérték a minőségi határukat.

Bibliográfia

1. George K. Zipf: *The Psychobiology of Language*. Houghton-Mifflin. (1935)
2. Cavnar, William B. and John M. Trenkle.: N-Gram-Based Text Categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (1994)
3. Füredi Mihály, Kelemen József: *A mai magyar nyelv szépprózai gyakorisági szótára*. Akadémiai Kiadó (1989)
4. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J.: *BLEU: a method for automatic evaluation of machine translation"*. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311-318. (2002)
5. Novák, Attila; László Tihanyi; Gábor Prószéky. *The MetaMorpho translation system*. In: Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, 111-114 (2008)
6. Bach Iván, Farkas Ernő, Naszódi Mátyás: *A magyar nyelv elemzése számítógéppel. Tervek egy természetes nyelvű interfészhez*: SzTAKI Tanulmány/199, ISBN: 963 311 230 3
7. Tomaž Erjavec: *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris, 2004.
8. Seidl. Péch Olívia: *Autentikus magyar szövegek és fordítás eredményeként létrejött célnyelvi magyar szövegek lexikai kohéziós mintázatának összehasonlító elemzése.* Tézisek 2011

Függelék: Nyelvi statisztikák az 1984 alapján

nyelv	állomány karakter	futó szó	szóalak	előfordulás>1 aránya	előfordulás>2 aránya	első 10 lefed hányad	első 100 lefed hányad	fele	Zipf S	Winston
orosz	497325	74105	18557	0,34714662	0,19787681	0,16714121	0,37602051	35	**0,786338**	619
magyar	576693	80643	**19233**	**0,34186034**	**0,19643321**	0,21508376	0,39130488	31	0,799111	**725/729**
litván	**493031**	**68485**	17387	0,35215966	0,20256513	0,12868511	0,33978243	42	0,812609	534
lengyel	559436	80191	19601	0,37763379	0,21621345	0,16806125	0,34962776	44	0,817449	562
török	553111	71867	19206	0,36707279	0,21639071	**0,11537979**	**0,28494302**	**68**	0,828763	839
észt	507108	72850	17007	0,35026753	0,20973716	0,17529169	0,37446808	32	0,831842	690
cseh	499161	79681	17630	0,37850255	0,22637549	0,15787954	0,36823082	34	0,863323	547/543
szerb	522198	87182	16557	0,39590505	0,23579150	0,20573054	0,42792090	20	0,889754	531
roszin	532257	89560	16873	0,39418005	0,23493154	0,20500223	0,42824921	20	0,891915	550
szlovén	538912	90164	16317	0,40019611	0,24250781	0,21406548	0,45360676	16	0,894747	**519**
bolgár	532488	85821	15227	0,40421619	0,24528797	0,21589121	0,45944465	14	0,897696	537
eszperantó	547610	89291	14477	0,3878566	0,23851626	0,23741474	0,48411374	11	0,917758	524
német	665606	99808	14060	0,40960170	0,25597439	0,17490581	0,49415878	10	0,97574	545
olasz	643510	102578	13771	0,45908067	0,29235349	0,20399110	0,47053949	12	1,026019	687
román	627065	107480	14155	0,45785941	0,29798657	0,20607554	0,47287867	12	1,029532	594/529
portugál	**678514**	**108376**	13824	0,45442708	0,29629629	0,20115154	0,47016867	12	1,047623	616
spanyol	566282	95380	11441	0,47111266	0,30705357	**0,26041098**	0,52680855	**77**	1,050474	790
francia	640977	103981	12247	0,47734139	0,31444435	0,21082697	0,51070868	91	1,065035	675
angol	587340	103740	**9304**	**0,53181427**	**0,36822871**	0,25881048	**0,54647194**	11	**1,182304**	527/515