Útravaló szép üzenet

Simon Eszter

2019. május 13.

1. Életbölcsességek

> „All dictionaries are incomplete, [...] there is no such thing as a perfect dictionary, there is, equally, no 'right' way to produce a dictionary."
>
> „the content and design of every aspect of a dictionary must, centrally, take account of who the users will be and what they will use the dictionary for"
>
> „no amount of theoretical rigour is worth a hill of beans if the average user of your dictionary can't understand the message you are trying to convey"
>
> „naked science is too delicate for the purposes of life"

2. Általános benyomások

Nyomtatott éra vs. elektronikus szótárak. A könyv szerzői még teljes mértékben a nyomtatott éra gyermekei, és nem is nagyon sikerült nekik az elektronikusba való váltás. A szótárszerkesztési szabályok lefektetése során egy könyv jelenhetett meg a lelki szemeik előtt, és valószínűleg azt vették figyelembe. A számítógépes eszközök, a korpuszok használata ugyan már a mindennapi rutin részévé vált, de még mindig a korpuszalapú (*corpus-based*) megközelítést követik. Ez a lexikográfiában azt jelenti, hogy az ember állítja össze a szótárba kerülő szavak listáját, határozza meg a szócikk felépítését stb., és ehhez segítségül hívja a korpuszokat. A másik megközelítés a korpuszvezérelt (*corpus-driven*), aminek alkalmazása azt jelenti, hogy a korpusz határozza meg, hogy mi fog szerepelni a szótárban és még akár azt is, hogy mi fog szerepelni egy szócikkben. Ebben a folyamatban automatikus módszerekkel épül a szótár a korpuszból, és az ember csak mint utószerkesztő szerepel. Az egymásnak megfelelő különböző nyelvű szavak összepárosításához, illetve az egyes szavak különböző jelentéseinek elkülönítéséhez használt legújabb módszer az ún. szóbeágyazásokon alapul. Ennek az a lényege, hogy a szöveg minden egyes szavához hozzárendelődik egy vektor, ami az adott szónak a különféle tulajdonságait reprezentálja, amelyek

között kiemelten fontos a szó kontextusa, vagyis az, hogy milyen más szavakkal szeret együtt járni. Az így kapott vektorokat aztán különféle metrikákkal össze lehet hasonlítani, és az elvárások alapján a leghasonlóbb jelentésű szavak vektorai lesznek a legközelebb egymáshoz.

3. Elméleti kérdések

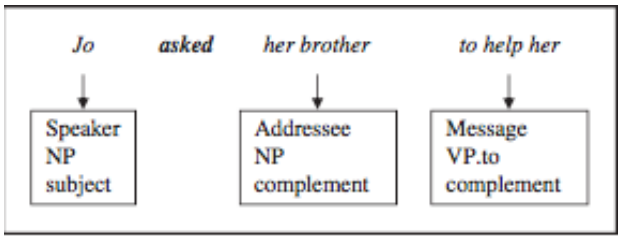
Metonímia. Más néven szabályos jelentésátvitel vagy még általánosabban szabályos referenciaátvitel. Azért szabályos, mert egy szemantikai mezőbe tartozó minden szóra lefuttatható ugyanaz a szabály. Például a CONTAINER FOR CONTENTS metonímia esetében minden CONTAINER szemantikai mezőbe és minden CONTENTS szemantikai mezőbe tartozó szóval meg lehet azt csinálni, hogy a CONTAINER-t mondjuk a CONTENTS helyett például ebben a mondatban: *Denise drank the bottle* (= the liquid in the bottle). És azért referenciaátvitel, mert mindezt tulajdonnevekkel is meg lehet csinálni – például: *Ted played Bach* (= the music of Bach → ARTIST FOR ARTFORM) –, és a tulajdonneveknek nincs jelentése, csak referenciája. A témáról bővebben lásd Simon [2013].

A könyv szabályos poliszémiának nevezi ezt a jelenséget, és csak később említi, hogy ez valójában a metonímia. Természetesen lehet poliszémiának is nevezni, mivel a helyettesítő szó (itt: CONTENTS, illetve ARTFORM) vonatkozhat saját magára és a helyettesítettre (itt: CONTAINER, illetve ARTIST) is. Lexikográfiai szempontból azért van ennek jelentősége, mert az egy szemantikai mezőbe tartozó szavakra egységes templátumot lehet kialakítani, illetve ezeknek a szavaknak a szerkesztésénél érdemes egységesen eljárni.

Metafora. A kognitív metaforaelmélet szerint a fogalmi metafora leképezést valósít meg egy forrástartományról egy céltartományra. A testesültség (*embodiment*) hipotézis [Lakoff and Johnson, 1980] szerint a forrástartomány mindig valami konkrét, fizikai, testi tapasztalatból származik, és a céltartományba tartozó absztrakt dolgokat ezeknek a segítségével tudjuk elsajátítani. Ami miatt ez a lexikográfia szempontjából érdekes lehet, az az, hogy amikor meg kell határozni egy szó elsődleges jelentését, akkor jól használható lehet ez a megközelítés.

Levin igeosztályai. Levin [1993] az igei vonzatkeret-alternációkat vizsgálva arra jut, hogy az ige jelentése befolyásolja az igei argumentumok realizációját. Az igeosztályok[1] két nagy csoportba oszthatók: vannak az alternációs osztályok (Chapters 1–8), ahol az egy osztályba tartozó igék ugyanolyan szintaktikai alternációt valósítanak meg, illetve a szemantikai igeosztályok (Verb Classes 9–57), ahol pedig jelentésük alapján csoportosulnak össze az igék. Ez utóbbiak esetében az osztály egy prototipikus tagjáról vannak elnevezve az osztályok: például a 30.1 osztályba tartoznak a „see" típusú igék, mint például a *detect*, a *hear*, a *notice* és maga a *see*. Ezek az igeosztályok jól használhatók a metonimikusan viselkedő tulajdonnevek felismerésére [Farkas et al., 2007]. Például ha

[1] `http://www-personal.umich.edu/~jlawler/levin.verbs`



1. ábra. Az ‘ask’ ige a REQUEST keretet hívja elő.

egy szervezetnév után a mondatban egy „say” osztályba tartozó ige következik, akkor ott nagy valószínűséggel egy ORGANIZATION FOR MEMBERS metonímiával van dolgunk, például: *IBM announced a forthcoming software capability.* Lexikográfiai hasznuk tekintetében hasonlítanak a metonímiára, lásd fentebb.

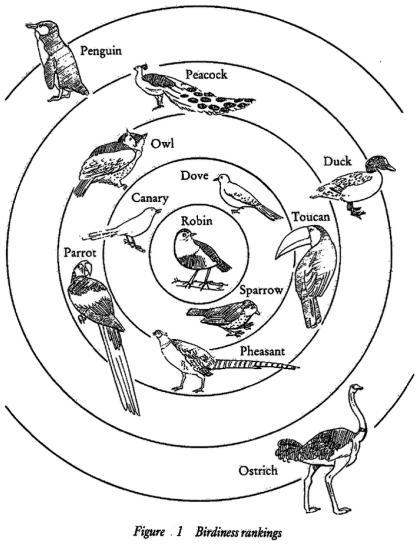
Keretszemantika (Frame semantics). Fillmore [2006] elméletében egy szemantikai keret egy szituációtípusnak a sematikus reprezentációját adja, annak tipikus résztvevőivel egyetemben – ez utóbbiak a keret elemei (*frame elements, FEs*). A keretszemantikán belüli elemzés során egy mondatban azonosítjuk a kötelező elemeket (*core elements*) és az opcionális elemeket (*peripheral elements*). Ezek nagyjából megfelelnek a vonzatoknak és a szabad határozóknak. Majd minden egyes elemhez hozzárendeljük a mondatban betöltött grammatikai funkcióját és tematikus szerepét, továbbá a frázis típusát. Az 1. ábrán arra láthatunk példát, ahogy az ‘ask’ ige előhívja a REQUEST keretet és annak elemeit.

A keretszemantika lexikográfiai haszna abban rejlik, hogy segítségével egyszerűen azonosíthatjuk azokat az elemeket, amelyeknek egy szó (jelen esetben ige) definíciójában mindenképpen szerepelnie kell. A keretszemantikát megvalósító erőforrás a FrameNet[2], ami ma már az angol mellett nyolc nyelven elérhető.

Arisztotelészi metafizika vs. prototípus-elmélet. Az egész nyugati világ gondolkodását rendkívül erősen befolyásolja az arisztotelészi metafizika, ami szerint a dolgok diszkrét kategóriákra oszthatók, és valami vagy beletartozik egy kategóriába, vagy nem. Eszerint minden dolog rendelkezik egy *differentia specifica*-val, ami mentén az a dolog elkülöníthető a többi dologtól. Ha a dolognak van ilyenje, akkor az X, ha nincs, akkor az nem X. Vagyis itt egy bináris döntésről van szó.

Ehhez képest az Eleanor Rosch nevéhez fűződő prototípus-elmélet [Rosch, 1973] azt mondja ki, hogy a dolgok egy bizonyos tulajdonság tekintetében egy skálán helyezhetők el, aminek az egyik végén a legtipikusabb dolog áll, a másik végén pedig az egyáltalán nem tipikus. A „madárság” az alappélda erre: a legtipikusabb madár az angolszász kultúrkörben a vörösbegy, míg a legkevésbé

[2] `https://framenet.icsi.berkeley.edu`



2. ábra. A madársági skála illusztrációja.

prototipikus a strucc meg a pingvin. A madársági skála illusztrációja a 2. ábrán látható.

Bizonyos területeken praktikusabb az utóbbi megközelítést alkalmazni: például az igekötők [Kalivoda, 2016] vagy a tulajdonnevek esetében [Simon, 2013] érdemes bizonyos tulajdonságokat megvizsgálni, amelyeknek bizonyos kombinációi arra utalnak, hogy a vizsgált dolog inkább ilyen, de kevésbé olyan.

Lexikográfiailag a jelentések elkülönítése és a definícióírás feladatában jön ez elő: egyes szavak definícióiban jellemzően megadnak egy magasabb rendű fogalmat (*genus*), ami alá besorolható a vizsgált dolog, plusz leírják azt a bizonyos *differentia specifica*-t, ami mentén elkülöníthető a *genus* fogalomtól. De vannak olyan esetek is, amikor ez nem alkalmazható, ilyenkor a prototípus-elméletnek jobban megfelelő módszer, ha adunk egy alap definíciót, majd olyanokkal specifikáljuk, hogy 'tipikusan' vagy 'jellemzően' ilyen vagy olyan szokott lenni.

4. Általános lexikográfiai kérdések

Terjedelem. Meg kell találni a középutat a fedés (*coverage*) és az elérhetőség (*accessibility*) között. Itt két, egymásnak ellentmondó igény merül fel: minél több szócikket ismertessünk a felhasználó számára minél könnyebben hozzáférhető módon. A könyv a nyomtatott érában íródott, és ebből a paradigmából nem is tud kilépni: a helyszűke indokol egy csomó döntést. Emiatt alakult ki az a gyakorlat is, hogy egy szócikkbe minél több információt akartak a szerkesztők belezsúfolni, de ma már szerencsére nem ez az általános – az angolszász lexikográfiában a 70-es évek óta inkább a felhasználóbarátság került előtérbe.

Felhasználók. A szótárépítés egyik legfontosabb szempontja a felhasználó: mindig azt kell szem előtt tartani, hogy kinek készül a szótár és milyen célra. Aktív (*encoding*) szótárra van szüksége azoknak, akik aktívan használják a nyelvet: írnak, beszélnek azon a nyelven, vagyis produkálniuk kell nyelvi szerkezeteket. Passzív (*decoding*) szótárat azok használnak, akik megérteni akarnak egy szöveget, de nem ismerik a benne levő szavakat, amihez elég a passzív tudás.

Mi tesz egy szótárat megbízhatóvá? A szótárnak az a feladata, hogy leírja a nyelvhasználatot, nem az, hogy előírja (*description vs. prescription*). Ehhez objektív, megbízható megfigyelésekre van szükség. Az introspekció nem elég objektív. A gyakori és elterjedt nyelvi elemeket kell vizsgálni, amit főleg korpuszokon tudunk megtenni.

A leírás vs. előírás kettőse mögött a racionalista és empirista nyelvfilozófiai megközelítés húzódik meg. A racionalista filozófiai tradíció Descartes és Leibniz nevéhez fűződik, akik szerint létezik egy velünk született nyelvi képesség, egy istentől származó, mindannyiunk fejében benne levő univerzális nyelvtan, amiből az következik, hogy egy grammatikai ítélet meghozatalához elég a fejünkbe nézni, és megtaláljuk a választ – ez lenne az introspekció. Ebben az esetben a grammatikalitási ítélet csak 1 vagy 0 lehet.

Az empirista filozófiai tradíció elsősorban Locke nevéhez fűződik, aki szerint tudásunk elsődleges forrása a tapasztalat; filozófiája az érzékszervi tapasztalat prioritásán alapul. Ebben az esetben tehát a grammatikalitási ítélet nem kétértékű, hanem fokozatai vannak. A nyelvre és a szótárírásra vonatkoztatva: az adatokra és azok gyakoriságára építve lehet megbízható állításokat tenni egyes szavakról.

Style Guide. A szótárkészítés alapja. Ebben lehet lefektetni a készülő szótárral kapcsolatos alapelveket, szabályokat, útmutatásokat. A Style Guide tartalma egy kontinuum, ami az általános alapelvektől az olyan specifikus szabályokig terjed, mint például hogy az etimológiai szótárra mindig 'TESz'-ként lesz hivatkozva. Ez nem egy kőbe vésett valami, hanem a szótárépítés során folyamatosan frissül az újabb és újabb kihívásoknak megfelelően. Az egyértelműen lefektetett szabályok és a kérdéses esetekre vonatkozó útmutatások nagy mértékben megkönnyítik és felgyorsítják a munkát, valamint biztosítják a konzisztenciát. Hasonló funkciót töltenek be a templátumok is, vagyis a hasonló tematikájú szócikkek generális szerkezetét felvázoló mintacikkek.

Szócikk. A 3. ábrán egy szócikkrészletet látunk. A szócikk más néven szótári bejegyzés (*dictionary entry*), ami lexikai egységekből (*lexical unit*) áll. A lexikai egységek az ábrán vastag számokkal vannak jelölve (1–5). A szócikk első eleme a címszó (*headword*), szintén vastaggal szedve. A lexikai egység nem összetévesztendő a lexikai elemmel, ami lehet szó, rövidítés, név, frázis, többszavas egység, bármi, amiről azt gondoljuk, hogy fel kell vennünk a szótárunkba és lexikográfiai leírást (definíciót vagy fordítást) kell hozzárendelnünk.

absolute / ˈæbsəluːt/ *adj* **1** complete or total: *I have absolute confidence in her.* | *We don't know with absolute certainty that the project will succeed.*
2 [*only before noun*] *especially BrE informal* used to emphasize your opinion about something or someone: *Some of the stuff on TV is absolute rubbish.*| *How did you do that? You're an absolute genius.* | *That meal last night cost an absolute fortune.*
3 definite and not likely to change: *We need absolute proof that he took the money.*
4 not restricted or limited: *an absolute monarch* | *Parents used to have absolute power over their children.*
5 true, correct, and not changing in any situation: *You have an absolute right to refuse medical treatment.*
LDOCE-4 (2003)

3. ábra. Egy szócikkrészlet.

Szótár vs. enciklopédia. A szótár nem enciklopédia, az enciklopédia pedig nem szótár. A szótárban a szavak jelentése van megadva (vagy a fordítása a két- vagy többnyelvű szótárak esetében), az enciklopédia pedig az adott címszóhoz kapcsolódó tényeket, ismereteket tartalmazza. Nem összekeverendő a Nagyszótár[3] és a Nagylexikon[4]!

5. Korpuszok

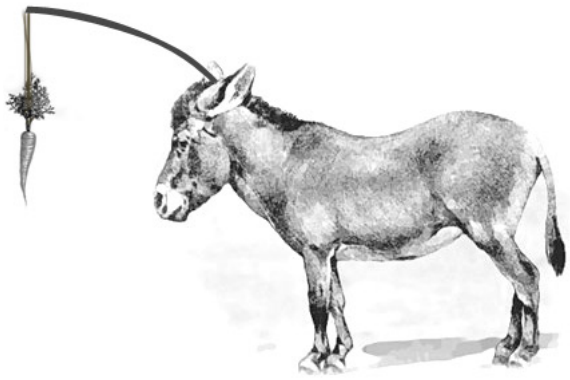
Egy jól használható korpuszdefiníció: „a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" [Sinclair, 2005].

A korpusz egy minta, a nyelvhasználatnak egy mintája. Nem a 'high quality' nyelvet kell reprezentálnia a korpusznak – egy lexikográfiai célú korpusznak inkluzívnak kell lennie, vagyis nem csak a „jó/helyes" nyelvi adatokat kell tartalmaznia.

A korpuszépítés egy „pragmatikus kihívás", ami jelen esetben azt jelenti, hogy a korpuszépítési álmok jellemzően módosulnak különféle praktikus és monetáris kényszerek hatására.

Egy készülő korpusz méretének meghatározásakor mindig figyelembe kell venni a Zipf-törvényt, ami szerint egy szó gyakorisága fordítottan arányos a

[3] `http://nagyszotar.nytud.hu/index.html`
[4] `https://hu.wikipedia.org/wiki/Magyar_nagylexikon`



4. ábra. A korpuszépítő és a reprezentativitás viszonya.

gyakorisági sorrendben betöltött helyével. Ebből az következik, hogy nagyon kevés szó adja ki a szöveg nagyon nagy részét. (Az angolban például a 100 leggyakoribb szó lefedi a BNC 45%-át.) A szótárépítés szempontjából ennek a fordítottja az igazán érdekes: a legtöbb szótári elem ritkán fordul elő. Ha nagyon sok szót tartalmazó szótárat szeretnénk építeni, akkor ahhoz nagyon-nagyon nagy korpuszt kell építeni, hogy még a legritkább szavakat is elégszer lássuk benne.

A soha el nem érhető cél egy reprezentatív korpusz építése lenne, amire törekedni érdemes, de azzal a kitétellel, hogy elérni nem lehet. A korpuszépítő és a reprezentativitás viszonyát a 4. ábra szemlélteti. Az igazi, elérhető cél egy kiegyensúlyozott (*balanced*) korpusz építése lehet, ami valahol a tökéletesen reprezentatív és a teljesen homogén, „monolitikus" korpusz között helyezkedik el. Egy ilyen korpusz törekszik arra, hogy a nyelv diverzitására reflektáljon, és lefedje a nyelvhasználat teljes repertoárját.

A könyv úgy csoportosít, hogy vannak a párhuzamos (*parallel*) korpuszok, amiken belül vannak a fordítási (*translation*) és az összevethető (*comparable*) korpuszok. Az NLP-ben nem ez a szokásos csoportosítás, hanem a két fő kategóriába a párhuzamos és az összevethető korpuszok tartoznak. A párhuzamos korpuszok ugyanannak a szövegnek két vagy több különböző nyelvű pontos fordítását tartalmazzák, mondatszinten illesztve. Az összevethető korpuszok esetében a két vagy több nyelvű szövegek közti hasonlóságot a téma, a származási idő vagy a származási hely adja. Például az egy eseményről angolul és magyarul közvetítő újságcikkek tekinthetőek összevethető szövegeknek.

6. Szoftverek

Sketch Engine. Egy európai projekt keretében szabadon elérhetővé tett lexikográfiai keretrendszer: `https://www.sketchengine.eu/`. Vannak benne előre betöltött és feldolgozott korpuszok, de lehet bele importálni saját korpuszt is.

Morfológiailag egyértelműsített korpuszt igényel, és ún. *word sketch*-eket gyárt saját nyelvtan alapján. Ezek a *sketch*-ek az egyes szavak lexikai profilját rajzolják ki, vagyis egy statisztikai alapú összesítést adnak a szó tipikus grammatikai és kollokációs viselkedéséről. A kollokációk erősségének méréséről lásd Kilgarriff et al. [2014]; Lexical Computing [2015].

Ezek a *sketch*-ek olyanokat is mondanak például, hogy egy adott főnév szívesen alanya bizonyos igéknek. Mindezt úgy, hogy jelenleg a *sketch*-ek előállítása a morfológiai szintű annotáción alapuló szabályokkal történik. Egy érdekes kutatási téma lehet, hogy hogyan lehetne a *sketch*-gyártást felturbózni szintaktikai (elsősorban dependencia) elemzéssel, illetve ha ilyet nem csináltak eddig, akkor miért nem.

A Sketch Engine tud még természetesen konkordancialistát is gyártani, ezt hívják másképp KWIC-nek (*KeyWord In Context*). A Sketch Engine (és szabadon felhasználható kistesója, a NoSketch Engine) lekérdezőnyelve (CQL) lehetőséget ad a korpuszokban található számos információ intelligens lekérdezésére.

Lexonomy Egy nyílt forráskódú, felhőalapú szótárszerkesztő keretrendszer: `https://www.lexonomy.eu/`. Szintén a Sketch Engine fejlesztőinek a munkája.

Hivatkozások

Farkas, R., Simon, E., Szarvas, Gy., and Varga, D. (2007). GYDER: maxent metonymy resolution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 161–164, Prague. Association for Computational Linguistics.

Fillmore, C. J. (2006). Frame semantics. In Geeraerts, D., editor, *Cognitive Linguistics: Basic Readings*, pages 373–400. Mouton de Gruyter, Berlin – New York.

Kalivoda, Á. (2016). A magyar igei komplexumok vizsgálata. MA Thesis, PPKE BTK, Elméleti Nyelvészet Tanszék, Budapest.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. Chicago University Press, London.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago.

Lexical Computing, L. (2015). *Statistics used in the Sketch Engine*.

Rosch, E. H. (1973). Natural Categories. *Cognitive Psychology*, 4(3):328 – 350.

Simon, E. (2013). *Approaches to Hungarian Named Entity Recognition.* PhD thesis, PhD School in Cognitive Sciences, Budapest University of Technology and Economics.

Sinclair, J. (2005). Corpus and Text – Basic Principles. In Wynne, M., editor, *Developing Linguistic Corpora: a Guide to Good Practice*, pages 1–16. Oxbow Books, Oxford.