SZTAKI Szótár - Olyan jó, hogy nem találom a szavakat

Pataki Balázs - pataki@sztaki.hu

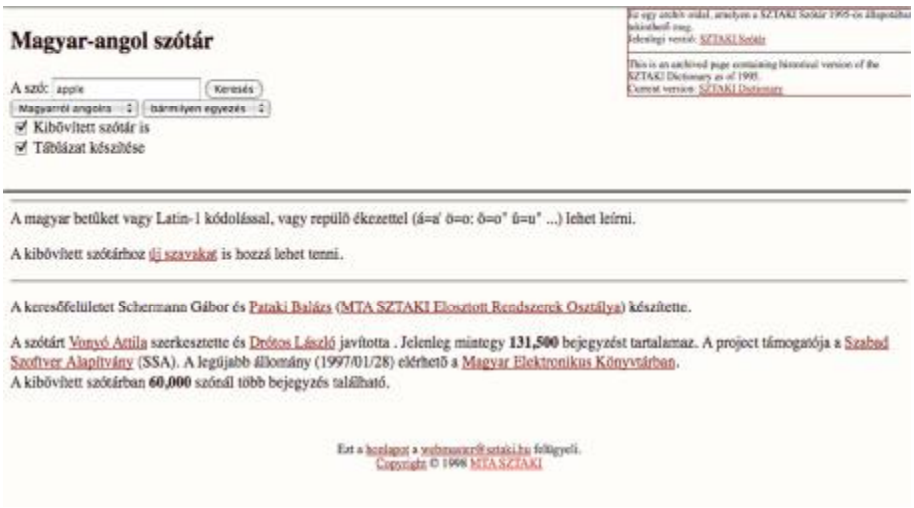
MTA SZTAKI - Elosztott Rendszerek Osztálya
1111, Budapest, Lágymányosi u. 11

A SZTAKI Szótár, a magyar web egyik legrégebbi és leglátogatottabb közhasznú szolgáltatása, 2012-ben mind külalakjában, mind szolgáltatásaiban teljesen megújult. Újragondoltuk a szótári adatszerkezeteket, lehetővé tettük, hogy elméletileg bármilyen nyelv bármilyen struktúrájú szócikkeit fel bírja dolgozni a rendszer, s mindehhez egy olyan szolgáltatás infrastruktúrát valósítottunk meg, ami lehetővé teszi olyan szótárak egyéni vagy közösségi fejlesztését, amik méretükben és minőségükben a papír szótárakkal vetekedhetnek. Az előadás keretében bemutatjuk az elért eredményeket, és a további terveinket.

A SZTAKI Szótár története

Képzeljünk el egy világot, amiben nincs se Google, se Amazon, se index.hu, se iPad, se szélessáv, és ezekre nincs is szükség. Ez az év 1995. Mi történt ekkor?

- Kevin Mitnicket letartóztatják mert betört az USA „legbiztonságosabb” szervereire
- AFC Ajax 1:0-ra legyőzte az AC Milant a Bajnokok Ligájában
- Dr. Ivan Turk megtalálja a világ legrégebbi sípját, amit a neandervölgyi ősember egy egész mamutból faragott ki.
- A moziban először látható egy egész estés 3d rajzfilm, a „Toy Story”, és ekkor rohangál Mel Gibson fel-alá a Skót felföldön, miközben az üvölti, hogy „aaaaaa” meg „grrrrrrr”
- Ebben az évben megy a Netscape a tőzsdére
- A Microsoft kiadja a Windows 95-öt, de jó üzleti érzékkel böngésző nélkül
- És ez az az év, amikor Walter Ostanek Grammy díjat nyer polka kategóriában
- .... valamint ennek az évnek a júliusában indul el a SZTAKI Szótár, ami tudomásunk szerint a magyar web első interaktív szolgáltatás



A SZTAKI Szótárat az MTA SZTAKI Elosztott Rendszerek Osztálya (DSD), mint a World Wide Web technológia egyik első hazai elkötelezett támogatója és kutatója fejleszti a kezdetek óta. A szótár fejlődésének szakaszai:

1995

- Megszületik a SZTAKI Szótár, az első magyar-angol internetes szótár.
- MEK ingyenes szókészletére épült (100.000 szó)
- Angol és magyar nyelvű keresőfelület
- Közösségi, bővíthetőségi funkció
- HTML 2.0, perl

1997

- Új német-magyar szótár
- MEK adatbázisára épült (40.000 szó)

1998

- Webster értelmező szótár és összekapcsolása az angol-magyar szótárral.

2000

- Új URL-eken is elérhető a szótár: szotar.sztaki.hu, dict.sztaki.hu
- Új szótár architektúra, HTML 4.0, PHP, dictd (szótári kiszolgáló szerver C-ben)

2001

- WAP-os felület (szótározás mobiltelefonról)
- Bookmarklet funkció
- Hibás szó jelzése a felületen keresztül
- Negyedévente frissülő vicces hírlevél a szótárról a nyitóoldalon

2002

- Új Francia-magyar szótár (9.000 szó)
- Kiejtési adatbázis, angol és német nyelvekre

2003

- Hasonlósági keresés, elgépelés esetén
- Nem pontos írásmód esetén, ragozott alakú szavak keresése
- Szótár egérmentesítése, billentyűvel elérhető funkciók
- Rövidítés adatbázis beágyazása
- CSS alapú dizájn

2004

- Szószedet: teljes szövegek vagy ottlapok kiszótározása
- Középsőujjas szótározás (böngészőbe épülő contextys alapú menük)
- Hangos szótár angolul, németül és magyarul

2006

- Hallatlan.hu integráció és mutogép
- Holland szótár 145.000 szópárral
- Olasz szótár (ma már) 190.000 szópárral
- Együttműködés az [origo]-val

2008

- MySQL alapú fulltext keresés

2010

- Az eddigi "szórt idős" fejlesztési tevékenységet felváltotta egy fejlesztési projekt, amihez a támogatást az MTA SZTAKI saját forrásból biztosította, miután az elmúlt évek alatt egyetlen pályázatot sem sikerült a projekthez elnyernünk.

A SZTAKI Szótár a megújulás előtt

A SZTAKI Szótár megújítása előttre az egyik legnagyobb forgalmú magyar weboldallá nőtte ki magát. A szolgáltatás naponta közel 1.000.000 keresést szolgál ki, több, mint 100.000 egyedi látogatónak. Ezt a forgalmat 3-5 kisebb nagyobb szerver szolgálta ki az idők folyamán.

A szolgáltatásnak sok évig integráns része volt az új szavak bevitelének lehetősége, majd később a talált hibáknak a jelzése. Ez egy idő után olyan méreteket öltött, hogy azt mi már nem tudtuk "szórt idős tevékenységként" magunk ellenőrizni, ezért a szavak hozzáadásának és javításának lehetőségét pár évvel ezelőtt leállítottuk.

Ez két dologot hozott magával. Egyrészt a szótáraink nem fejlődtek a továbbiakban olyan mértékben, ahogy azt szerettük volna. Ez azt jelentette, hogy sok hiba javítatlanul maradt, új szavak nem (nagyon) kerültek az adatbázisba, illetve, mivel a kiinduló adatbázisainkból is hiányoztak az alapvető nyelvtani jelzetek, szavak szótári alakjai, stb. ezért a nyelvtanulók számára nem minden esetben voltak kielégítők a SZTAKI Szótár által megjelenített eredmények.

Kiszolgálás szempontjából viszont így némileg egyszerűbb feladatunk volt, mivel egy nem, vagy alig változó adatbázist kellett kiszolgálnunk, amihez hatékony cachelési eljárások voltak segítségünkre.

Az "Újszótár" projekt keretében azonban éppen azt szerettük volna, ha szótáraink organikusan fejlődő, igazi közösségi térként is működő entitásokká válnak, ami előrevetítette, hogy ennek kiszolgálási infrastruktúrja jóval több erőforrást igényel, mint ami eddig rendelkezésünkre állt.

Az "Újszótár" szolgáltatásai

A SZTAKI Szótár megújításával az alapvető céljaink a következők voltak:

- *A SZTAKI Szótár a szótárkészítés magyar (sőt európai) platformjává váljon.* Főleg a kis nyelveket beszélő országok problémája, hogy igen nehéz modern szótárakhoz hozzájutni, mivel azok a kiadók számára általában nem rentábilisak. Jó példa erre a norvég-magyar szótár, amiből csak egy nagyon régi volt elérhető, és a jogtulajdonos annak internetes publikálását sem engedélyezte, ezért érdeklődők önszerveződő módon létrehozták a http://dict.hunnor.net/ oldalt, amin keresztül közösségileg építenek egy szabadon hozzáférhető norvég-magyar szótárat. A SZTAKI Szótár az ilyen kezdeményezéseket szeretné összefogni, és infrastruktúrát nyújtani, hogy bármilyen nyelvpárral tetszőleges átfogású két- vagy egynyelvű szótárak épülhessenek.

- *A közösségi szótárkészítést támogató eszközrendszer valósuljon meg*. Az "Újszótár" alapja egy közösségi szótár szerkesztő rendszer, amivel új

szótárakat lehet létrehozni, szócikkeket szerkeszteni, azokhoz új adatokat (jelentéseket, nyelvtani alakokat, stb) hozzáadni, a hibákat jelezni és javítani lehet. Mindezt egy olyan workflow támogatással, aminek a segítségével a szótárak szerkesztői kioszthatják, illetve elvállalhatják az egyes feladatokat, s közben módjuk van az egyes javítási javaslatokat megvitatni és értékelni is. A szótárak szerkesztését támogató közösségi eszközökön túl célunk volt az is, hogy hagyományos közösségi eszközök, pl. a szótárakhoz kapcsolódó blog, és fórumok is megjelenjenek.

- *Új szótári adatszerkezet.* Az eddigi évek tapasztalatai, valamint a megismert klasszikus szótárak alapján olyan új adatszerkezetet szerettünk volna megvalósítani, ami lehetővé teszi tetszőleges nyelvek szótárainak szócikk szerkezetének ábrázolását, beleértve a szavak szótári alakjainak, a különböző osztályozási jelzetek, valamint egyéb társított információk felvételét.

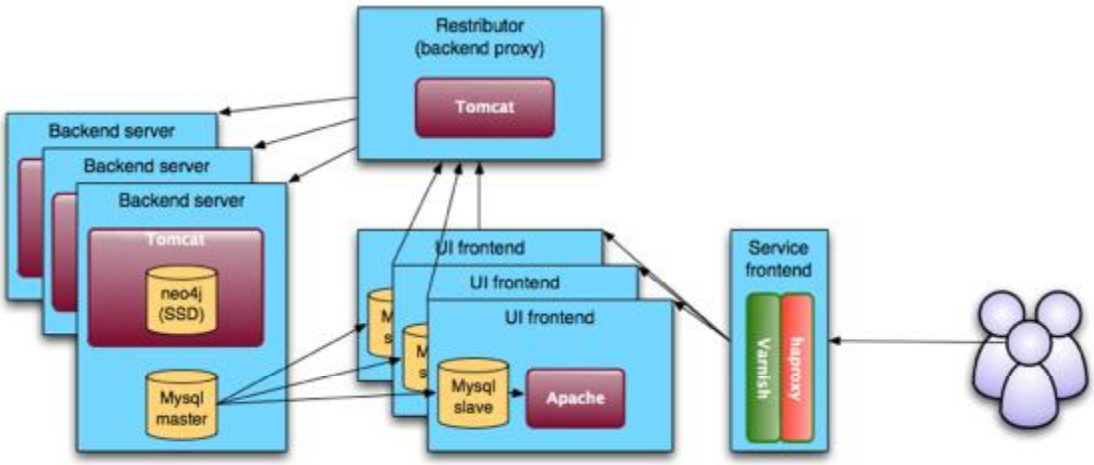
- *Az eddigieknél hatékonyabb, ha lehet még gyorsabb keresés az adatbázisban.* A SZTAKI Szótár egyik legnagyobb előnye volt mindig is, hogy nagyon gyorsan lehetett benne szavakat megtalálni. A hozzáértőbb felhasználók ezenkívűl beállíthatták a keresés módját, valamint egyéb paramétereit. A megújult szótárban olyan megoldást szerettünk volna, ami lehetőség szerint zéró beállítással is azonnal és nagy valószínűség szerint a megfelelő találatokat adja a fealhasználóknak, illetve hogy a beállításokat ne kelljen előre beállítani (pl. a fordítás nyelvét és irányát, vagy a keresési módot), hanem azt a keresés közben az eredmények ismeretében, mint szűrő feltételt lehessen megadni.

- *Esztétikus új felhasználói felület a fenti célok kiszolgálása érdekében.* A szótár belsejének megújításával párhuzamosan olyan külsőt is szerettünk volna kölcsönözni a szótárnak, aminek segítségével minden platformon kényelmesen és hatékonyan érhetők el a szolgáltatás funkciói. Itt kiemelt fontosságú volt, hogy a desktop eszközökön túl a mobil készülékeken (okostelefonokon, tableteken) is használható legyen ugyanaz, vagy csak minimálisan átalakított felhasználói felület.

Az "Újszótár" architektúrája

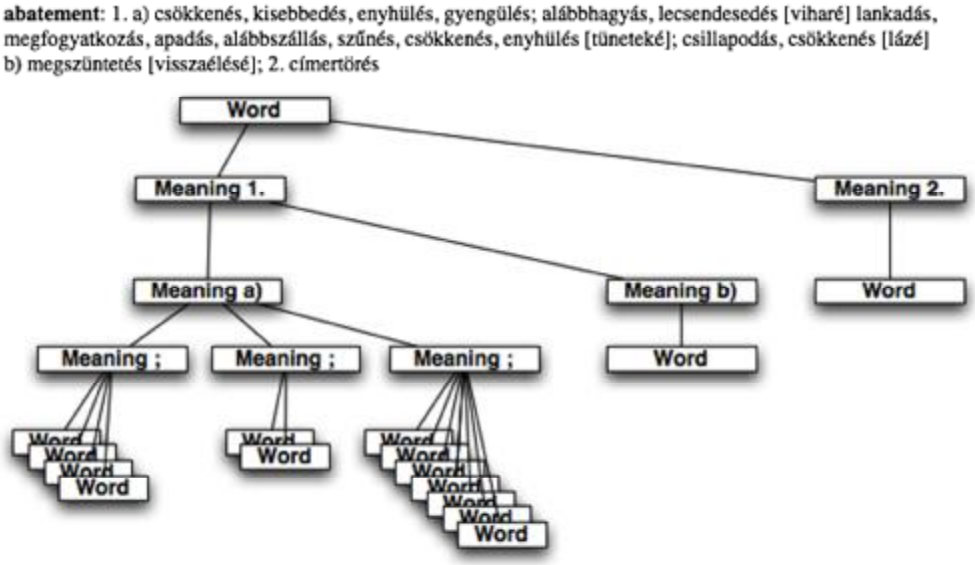
A SZTAKI "Újszótár" két nagy funkcionális egységből áll:

- Szótári adatbázis kezelő és kereső (backend), aminek egyes példányai egy elosztó proxy szerveren (restributor) keresztül érhetők el
- UI frontend közösségi funkciókkal (frontend), mely előtt előtétként terheléselosztó és caching szerverek (Varnish, haproxy) találhatók



Szótári adatbázis backend

Az "újszótár" backendjének alapját egy olyan hálós, nem SQL alapú adatszerkezet biztosítja, amit a neo4j adatbáziskezelő segítségével valósítottunk meg. Ebben a rendszerben a szótárak és szócikkek objektumai gráf csomópontokként, kapcsolataik élekként vannak tárolva. Ebben az adatszerkezetben egyes csomópontok több szótárban is használhatók, vagyis ugyanaz a vezérszó, vagy ugyanaz a "jelentés" csomópont több szótárban is használható, így a megfelelően összekötött csomópontokon keresztül a jövőben lehetőség lesz olyan bejárásokat is végeznünk, amelyek segítségével jó minőségű keresztszótárakat is létre tudunk hozni.



A fenti ábrán egy klasszikus szócikk és annak SZTAKI Szótárbeli gráf ábrázolása látható. Egy vezérszóhoz több jelentés, és aljelentés tartozhat, míg végül a jelentés fa alsó csomópontjaihoz kapcsolódnak az adott jelentésben egyenértékű szinonímák.

Azonban a jelentések értelmezéséhez csupán a kapcsolatos megléte nem elegendő, azok finomítását minden szótárban kiterjedt jelzet rendszer is segíti. Ezekkel a jelzetekkel adhatók meg a szavak vagy jelentések stilisztikai értéke (régies, irodalmi, szleng), tematikus besorolása (katonai, oktatási, sport), földrajzi felhasználási területe (brit angol, amerikai angol), etimológiai magyarázata, stb. A neo4j adatbázisban mind

a csomópontok, mind az élek rendelkezhetnek tulajdonságokkal, s ezen tulajdonságok hordozzák a SZATKI Szótárban használt kiterjedt jelzetrendszert.

Szótáranként, de tipikusan nyelvenként eltérő lehet, hogy mit jelent egy szó "szótári alakja", milyen állandó attribútumokkal, kapcsolatokkal kell rendelkezni a vezérszónak, s milyen jelzetekkel a jelentéseknek. Ehhez szintén gráf struktúrában megadható szótáranként a szócikkek szerkezete. Ez a séma határozza meg a szócikk szerkesztő számára, hogy milyen csomópontokat lehet vagy kell felvenni, milyen kötelező vagy szabadon megadható attribútumokkal, s az egyes csomópontok között milyen kapcsolatokat kell létesíetni. De ugyanezt a sémát használjuk arra is, hogy a keresési eredményeket kigeneráljuk a felhasználóknak. Ugyanahhoz szótárhoz akár több ilyen séma is tartozhat, így a kigenerált adattartalom akár a céleszközre is szabható, vagyis desktop eszközön több és bővebb adatot szolgáltathatunk, míg mobil eszközökre csak egy leegyszerűsített verziót.

Szótári felhasználói felület frontend

Az "Újszótár" felhasználói felületét Drupal alapon valósítottuk meg. A Drupalt azért választottuk, mert alapból sok olyan funkciót és modult ad, amikkel a közösségi szolgáltatásaink (blog, fórum, jogosultság kezelés, stb) egyszerűen lefedhetők voltak.

A felhasználói felületet két részre bontottuk: a keresőre és a közösségi funkciókra.

A kereső a SZTAKI Szótár legfontosabb és leginkább használt funkcionalitása, ezért erre különös gondot fordítottunk. A célunk az volt, hogy egy olyan kényelmes HTML5 alapon működő keresőt valósítsunk meg, ami előzetes beállítások nélkül is azonnal és nagy valószínűség szerint a megfelelő találatokat adja a felhasználóknak.

A "régi" SZTAKI Szótárban a keresési feltételeket mindig előre be kellett állítani (milyen nyelvről, milyen nyelvre, milyen keresési móddal, stb). Mára a felhasználók annyira túlterhelődtek választási lehetőségekkel a webes szolgáltatásokat használva, hogy a sok beállítási lehetőségtől megijednek, vagy egyszerűen figyelmenkívül hagyják őket. Ha pedig így van, akkor csak az lehet célunk, hogy ne is kényszerítsük őket választásra és beállításra előre, viszont, amint már van pozitív visszajelzése (van

eredménye a keresésének), akkor tegyük lehetővé a találati lista szűkítését "rávezető" segítségekkel.

A fentiek figyelembevételével a keresőfelület középpontjában a kereső mező van, amivel alapból bármilyen nyelven kereshetők bármilyen nyelvi fordítások. Ha elkezdjük begépelni a keresőszót, alatta "autocomplete" módon azonnal megjelennek az illeszkedő vezérszavak, illetve egy rövid kivonat, hogy egyes nyelveken milyen fordításai érhetők el a vezérszónak. További rávezető segítségként azt is kijelzi a kereső, hogy az adott keresőszó milyen nyelven értelmezhető egyáltalán. Így, már keresés közben lehet finomítani a keresési feltételeken.

Hasonló módon, ha az "autocomplete" listából kiválasztunk egy szót, és megjelennek annak fordításai, akkor ismét csak a találati lista alapján szűkíthetünk például kereső nyelvre, nyelvpárra, vagy szótárra.

Azon kívül, hogy a felhasználók nem szeretnek döntéseket hozni - pláne előzetesen nem - azt sem kedvelik, ha olyan információkkal terheljük őket, amikre nincsen szükségük. Azt, hogy egy szótári szócikkből kinek éppen milyen információra van szüksége lehetetlen előre kitalálni, viszont segíthetünk azzal, hogy különböző részletességű találati listákat generálunk. Az "Újszótárban" alapból a teljes nézetet adjuk, ahol az összes lehetséges szócikk tartozék megtekinthető, de ez "visszabutítható" a klasszikus SZTAKI Szótári nézetre, ami a "Villámnézet" elnevezést kapta. Ezen kívül jelenleg kísérleti jelleggel megtalálható egy "Papír" nézet is, ami táblázat helyett a papírszótárakból ismert szekvenciális megjelenítést produkálja. Mindez a nézetváltás azonnali módon, az oldal újratöltése nélkül állítható, így a keresési és böngészési élményt ez sem rontja.

A találati listák szócikkei azonban nem pusztán az eredmények megjelenítésére szolgálnak, hanem átvezetnek a közösségi funkciókhoz is, ugyanis a szócikkekben "in situ" módon lehetséges javításokat javasolni, új szavakat felvinni, vagy egy létező szócikkhez új fordításokat javasolni.

Ezek a javaslatok aztán bekerülnek az egyes szótárakhoz tartozó workflowkba, amiket a szótárak szerkesztői elbírának. Az egyes szótárak szócikkeihez alapból bárki tehet javaslatokat (de ez szótáranként akár letilható), de még jobb, ha valaki a szótári közösség tagjává válik a "csatlakozom" gomb megnyomásával. Ezáltal részt vehet a szótárnak, mint közösségnek a munkájában, jogosulttá válhat, hogy a szótár szerkesztője legyen, írjon a blogba, részt vegyen a fórum beszélgetésekben.

Az "Újszótár" a felhőben

Mint korábban említettük, a megújult SZTAKI Szótárnak jelentősen más lesz a terhelési karakterisztikája, mint volt a régen. A szótári adatbázisok dinamikusan fognak változni, a frontend Drupal alapú portáljának oldalai csak részben, vagy egyáltalán nem cachelhetők, stb. Mindezen okból a jelenlegi szolgáltatási infrastruktúránkat is át kellett gondolni, s a SZTAKI "Újszótár" éles üzeme csak ezen új infrastruktúra üzembeállítása után lesz lehetséges. A jelenlegi elképzeléseink szerint a SZTAKI Szótárat cloud alapokra fogjuk helyezni, így biztosítva, hogy növekvő erőforrás igények esetén is zökkenőmentesen ki tudjuk szolgálni a felhasználóinkat. jelenleg zajlik a SZTAKI saját cloud infrastruktúrájának tervezése és kivitelezése, aminek első felhasználója a megújult SZTAKI Szótár lesz.

Összefoglalás

2012-re elkészült a SZTAKI "Újszótár", ami a 17 éve töretlen népszerűségű SZTAKI Szótár kívül és belül is megújított jövőálló szolgáltatása. Az "Újszótár" célja, hogy a szótárak készítését és keresését mindenki számára elérhetővé tegye, s olyan szótár készítési platformmá váljon, aminek segítségével a papír szótárakkal megegyező minőségű és alaposságú online szótárak készíthetők és publikálhatók. Azt szeretnénk, ha a szótárazóink a szolgáltatást használva azt mondanák *"SZTAKI Szótár - olyan jó, hogy nem találom a szavakat .... de majd akkor én hozzáadom azokat"*.