

4. A SZOFTVEREKRŐL

4.1. Bevezetés

Napjainkban mindenki a saját bőréen tapasztalhatja, hogy a túlzott információbőség nemhogy segítené a világban való tájékozódást és a világ megértését, hanem inkább elbizonytalanít és esetleg a káosz érzetét is keltheti bennünk. Különösen igaz ez akkor, ha az információ rendezetlen és zuhatagként önt el bennünket. A korpuszok is hatalmas mennyiségű információt tartalmaznak, nemcsak nyelvészeti szempontból. Így tehát az információlekérdezés pontossága, gyorsasága és minősége kulcsszerepet játszik a korpuszok használatával elérhetővé vált információk elemzésében és értelmezésében, ami a használt szoftverek tulajdonságaitól nagymértékben függ.

Az előző fejezetekben többször említettük, hogy az annotált korpuszok jelentős hányadának esetében a korpusz használatához külön szoftvert fejlesztettek ki, amit csak az adott a korpuszsal lehet használni. Könnyen belátható, hogy egy bizonyos programot nem lehet két különböző annotációval rendelkező korpusz esetében eredményesen használni, hacsak a program egy részét át nem írják. Ebben a fejezetben olyan programokat igyekszünk bemutatni, amelyek könnyen hozzáférhetőek magánszemélyek számára is. Mivel ezek között magyar nyelvű nem található, azt a megoldást választottuk, hogy az idegen nyelvű program menüit és használatát képekkel illusztrálva részletesen leírjuk. Ezzel szinte egy magyar nyelvű használati utasítást nyújtunk, amit akár más, hasonló program esetében is használhat az olvasó. A legnépszerűbb (és legjobban használható) „fizetős” programok is említésre kerülnek, de elsősorban az internetről ingyen letölthető programokról esik majd szó. Így anyagi lehetőségeitől és nyelvi igényeitől függően választhat az olvasó, hogy melyeket szeretné kipróbálni. A megvásárolandó programok korábbi változatai is sokszor ingyenesen letölthetők, sőt az új verziók bizonyos ideig, általában 2-4 hétig, ha megkötésekkel is, de szabadon használhatók. Javasoljuk, hogy a könyvben szereplő programokat lehetőleg az olvasással egy időben próbálja ki az olvasó.

4.2. A korpuszok készítésekor használt programok

Más eszközökre van szükség és más eszközök használhatók eredményesen, ha a folyóból magunk akarjuk kifogni a halat a vacsorához, vagy ha a boltban élőhalként vesszük, vagy ha félkész mélyhűtött áruként. Így más programokra van szükségünk, ha a használandó korpuszt teljesen magunknak kell elkészíteni, vagy ha „félkész” korpuszsal dolgozunk. Még egy hasonlóság a horgászattal az, hogy ha magunk fogjuk a halat, akkor azt esszük, amit a jószerecske a horogra akasztott, és nem válogathatunk túl sokat,

különösen, ha időre kell a vacsorát elkészíteni. A félkész termékek között azonban válogathatunk, és valószínűleg gyorsabb lesz a vásárlás, mint a türelmes pecázás. Végül még egy érv szólhat a félkész termék mellett: sokan irtóznak vagy nem is tudják, hogy hogyan kell halat pucolni. A korpuzkészítést is ajánlott a korpuz használatának jobb megismerése utánra halasztani. Így ebben a fejezetben a korpuzokban rejlő információ lekérdezésére használt programokat ismertetem.

A „nyers” korpuzból a „félkész”, fogyasztó számára is megfelelő korpuz elkészítéséhez vezető úton általában a következők történnek – annak ellenére, hogy a korpuz annotációról már az előzőkben szóltunk, szükségesnek érezzük, hogy e fejezet elején dióhéjban felelevenítsük a legfontosabb tudnivalókat. Egészen néhány évvel ezelőttig alapvető feltétel volt, hogy a korpuz csak szövegfájlokból állhatott, amelyek kiterjesztése txt volt. A szövegfájlon belül természetesen nemcsak a szöveg szerepelt, hanem az arra vonatkozó információ is. A szövegre vonatkozó információt a számítógép számára is olvasható módon meg kellett különböztetni a szövegtől. A legegyszerűbb esetben a fájl elején szerepelt a szöveg eredetére vonatkozó információ, és ezen kívül csak a paragrafusokat jelölték. Ezen esetekben a keresés a szöveg egyes elemeire vagy írásjelekre korlátozódott. Ez még meglehetősen „nyers” korpuz, amit viszonylag egyszerűen magunk is létrehozhatunk.

Ha valaki jártas a honlapkészítésben, akkor jól tudja, hogy a honlapon látni kívánt szövegeket kódok veszik közre, amelyek a megjelenítésre vonatkozó információt tartalmazták, de ezek a honlapon nem jelennek meg. Ehhez hasonló az annotáció jelölése is, ahol ilyen jelek fogják közre a szövegre vonatkozó információt.

A szófaji azonosítás (tagging) esetén minden egyes szövegszót címkével látnak el, és ezeket a címkéket „visszaírják” a szövegbe. Természetesen ezt kézzel is el lehet végezni, ha van rá néhány évtizedünk. Ezt a munkát azonban egy címkéző programmal, azaz taggerrel végzik. A *tagger* előzetes nyelvi elemzések és szólisták (szótárak) eredményei alapján készítik, és még „nem látott” szövegeken tesztelik, hogy a hibaszázalék minél kisebb legyen. Mivel 100%-os pontosságú program nincs, ezért vagy kézzel ellenőrzik, vagy tudomásul veszik, hogy „vannak benne hibák”. A szófajilag annotált szövegben már nemcsak szövegszókra kereshetünk, hanem a homográfok (azonos írásképi, de különböző jelentésű szavak) esetében megadhatjuk, hogy milyen szófajt keresünk, például *ég*, főnévként vagy igeként. Vagy kikereshetjük az összes melléknevet anélkül, hogy találgatnunk kellene, vajon például a *tündéri*, és a *mesés* szerepel-e egy szövegben.

A morfológiailag bonyolult nyelvek esetében, mint például a magyar vagy japán, a szófaji elemzésen kívül a morfológiai elemzésre is nagy szükség van. Míg az angolban a lehetőséget külön szóval fejezik ki (*I can go home now.*), és így akár egy nyers szövegben is viszonylag könnyen kikereshető, a magyar nyelv esetében (*Hazamehetek.*), ha a *-hat/-het* kifejezésre keresnénk, rengeteg más szó is szerepelne a listánkon, például *hetes*, *heten*, *hatvan*, *hetven*, csak hogy néhányat említsünk. A morfológiai elemző programok is már előzetes nyelvi elemzésekre épülnek. A magyar helyesírást és nyelvtant elemző program részét képezi egy morfológiai elemző program, melyet a Morpho-Logic nevű magyar cég készített. A morfológiai elemző programokkal nemigen talál-

kozhat önmagában a nagyközönség. A morfológiai elemzés eredménye is „visszakerül” a korpuszba.

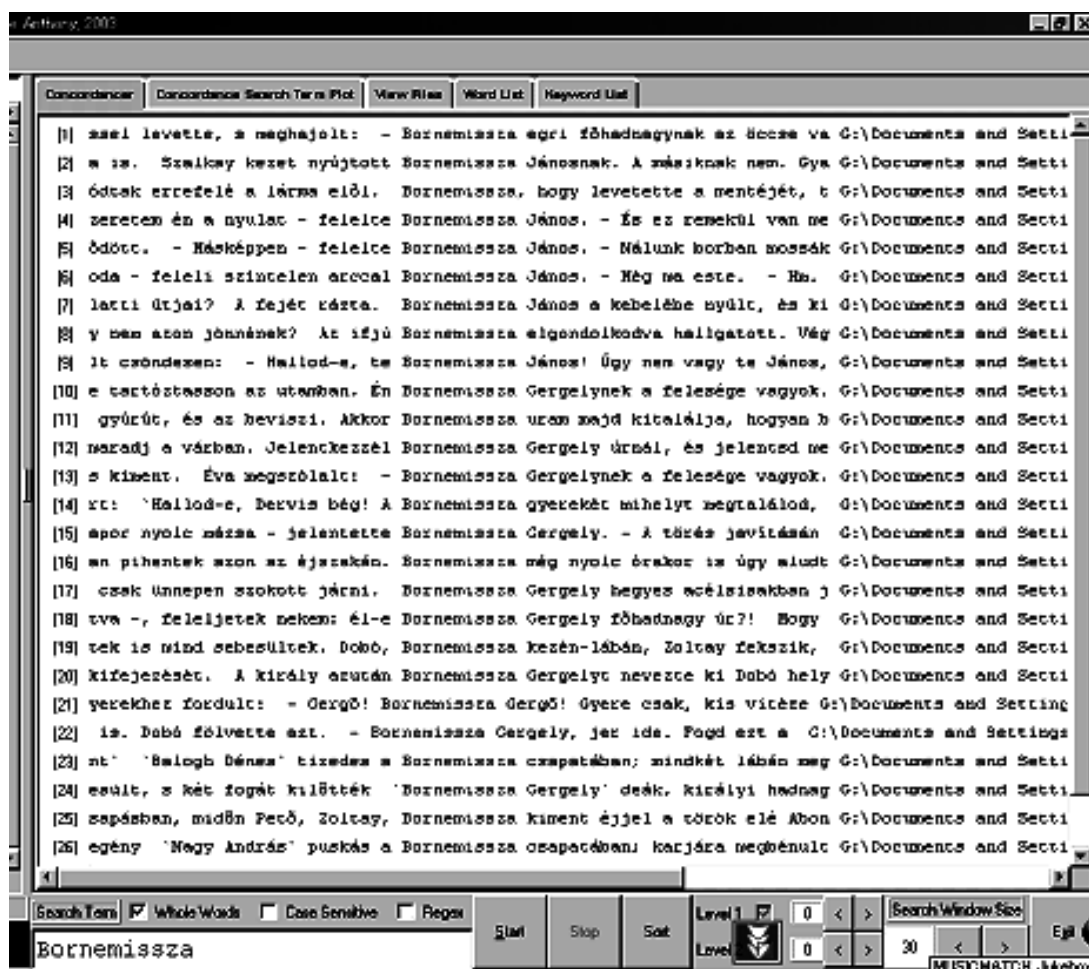
A szintaktikai elemzés az angolban megint csak viszonylag egyszerű, hiszen a szórend viszonylagosan kötött. A kötetlen szórend esetében viszont a morfológia általában segít vagy esetleg egyértelműen meg is határozza a szintaktikai kapcsolatokat. Ezen elemzések eredményei is visszakerülnek a korpuszba. Így tehát a *Tejet ivott lefekvéskor* mondatban szereplő *tejet* szó mellett a korpuszban az az információ is fel lesz tüntetve, hogy ebben a mondatban tárgyként szerepel.

Szó esett már olyan korpuszról is, amely angolul tanuló diákok írásaiból áll. Ebben a korpuszban a nyelvtanár a diákok hibáit és azok fajtáit látta el kódokkal. A kódok lehetővé teszik, hogy bizonyos típusú hibákra keressünk a korpuszban, vagy egyszerűen szám szerint összehasonlítsuk a különböző fajtájú hibákat. Az annotáció variációi szinte korlátlanok, így bárki bármilyen kódot kitalálhat a saját szükségleteinek megfelelően.

Minden olyan információra, amely a korpuszban fel van tüntetve, viszonylag egyszerű programmal is rá lehet keresni. Ebből következik, hogy a már annotált korpusz használata esetén sokkal pontosabb és gyorsabb lesz a keresés és lekérdezés, mint ha csak pusztán szövegben keressünk. Viszont a korpusz előkészítése sok időt, energiát, és ha nem kézzel végezzük, speciális programokat igényel. A következőkben az információt lekérdező programokról esik majd elsősorban szó.

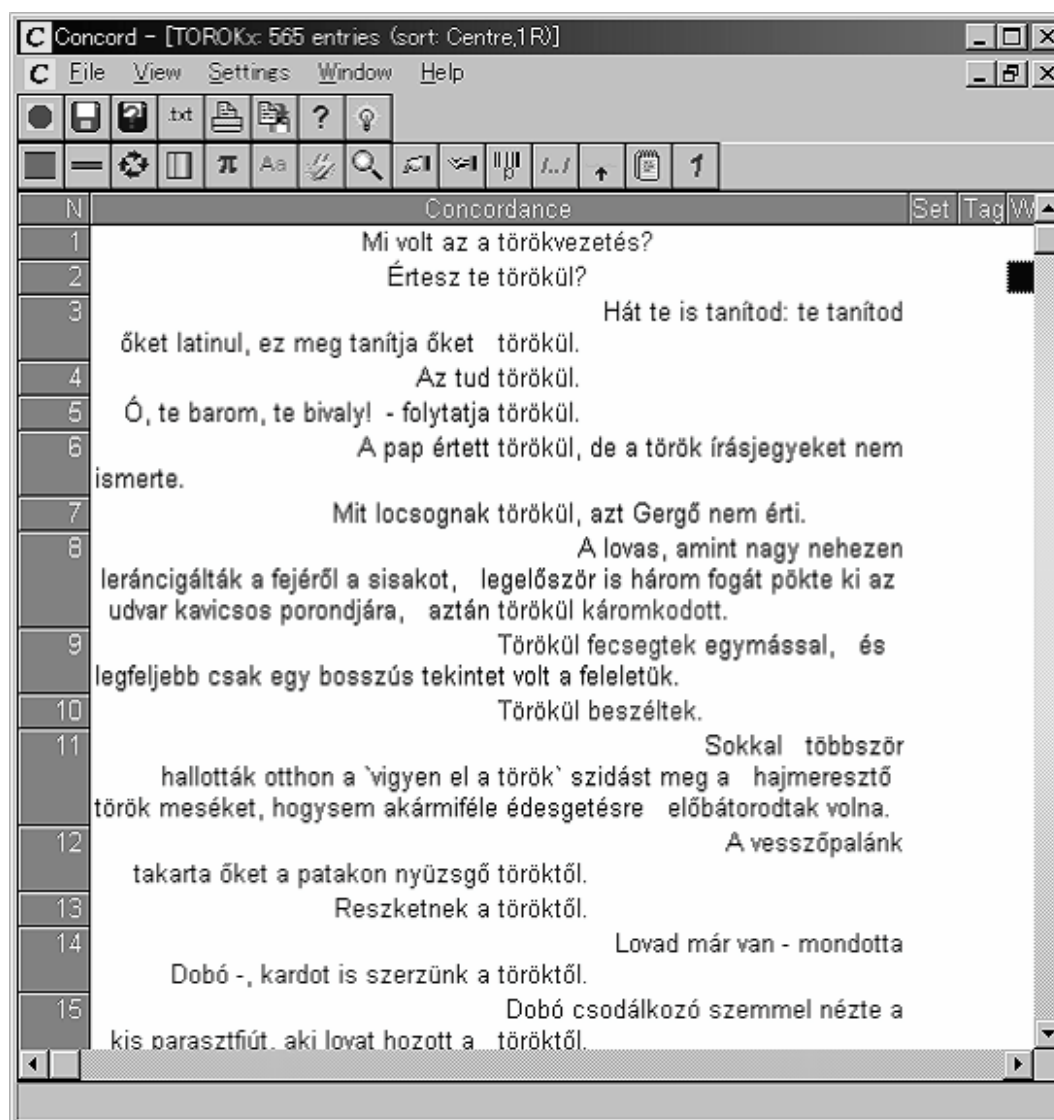
4.3. A konkordanciaprogramok

Bizonyára mindenki, aki használ szövegszerkesztőt, került már olyan helyzetbe, hogy a már elkészített és meglehetősen hosszú szövegben valamit ki kellett javítania, vagy hozzá kellett tennie valamit a már leírtakhoz, de ezt a megfelelő helyen kellett megtennie. Erre valószínű, hogy a Szerkesztés menü Keresés almenüjét használta, melynek segítségével gyorsabban megtalálta a kérdéses pontot. Ez esetben a keresett szó első, második stb. előfordulását könnyen meg is találhatta, de egyszerre csak egy előfordulást lehetett látni. A konkordanciaprogramok abban különböznek ettől a funkciótól, hogy nemcsak kikeresik a keresett elemet, hanem az elem összes előfordulását a szöveggörnyezettel együtt „kimásolják” egy külön ablakba. Így lehetővé válik, hogy egyszerre tekintsük meg a keresett elem előfordulásait a szöveggörnyezettel együtt. A vizuális elrendezés is segíti a gyors felismerést, hiszen a keresett elem mindig a képernyő közepén, azonos helyre kerül. A szöveggörnyezet általában nem teljes mondat, hanem csak annyit mutat egyszerre, amennyi a képernyőre a keresett elem előtt és után kifér. Ez lehet több, mint egy mondat, vagy csak egy mondatrészlet. Ezt a megjelenítési formát szokás KWIC, azaz Key Word In Context, magyarul a „kontextusban levő kulcsszó” formának nevezni.



29. ábra: „Bornermissza” konkordanciája az Egri csillagokból (AntConc program)

Ha valakit zavar, hogy nem teljes mondatokat lát a képernyőn, akkor két lehetőség közül választhat. Vagy saját kezűleg törli ki a mondatfördékeket és egészíti ki a hiányzó részeket, vagy olyan programot választ, amely arra is képes, hogy egész mondatokat tegyen láthatóvá a képernyőn. Ebben az esetben azonban előfordulhat, hogy a teljes mondat két vagy több sort foglal el, a mondat hosszától függően. A következő ábra ezt a megjelenítési módot szemlélteti. A 8. és 11. példamondat esetében azt is megfigyelhetjük, hogy ezek három sort foglalnak el.



30. ábra: Konkordanciák mondat formában (WordSmith program)

Jóllehet a konkordanciaprogramok erről a funkcióról kapták nevüket, manapság a legtöbb ilyen program számos más funkciót is magában foglal, így nem csupán konkordanciák készítésére alkalmasak, hanem a szövegre és a keresett szóra vagy kifejezésre vonatkozó alapvető statisztikai információkkal is szolgálnak. A szövegszerkesztők, mint például az MS Word, is képesek a teljes szövegben szereplő összes szót megszámlálni, de arra már nem képesek, hogy listát is adjanak a szövegben előforduló összes szóról (azaz típusokról, angolul: *types*) és az egyes szavak előfordulásának gyakoriságáról. Az alábbi ábra gyakorisági sorrendben mutatja a szövegben szereplő szavakat. A szó mellett az előfordulások száma és a szöveghez mért százalékos arányuk látható.

N	Word	Freq.	%	Lemmas
1	A	5,839	11.20	
2	AZ	1,541	2.96	
3	ÉS	783	1.50	
4	HOGY	672	1.29	
5	IS	608	1.17	
6	NEM	575	1.10	
7	S	556	1.07	
8	EGY	528	1.01	
9	MEG	404	0.77	
10	TÖRÖK	391	0.75	
11	DOBÓ	340	0.65	
12	CSAK	315	0.60	
13	VOLT	295	0.57	
14	DE	292	0.56	
15	MÁR	233	0.45	

31. ábra: Gyakoriság szerinti szólista az *Egri csillagokból* (WordSmith program)

A szólista minden egyes alakot külön vesz, így ha úgy érezzük, hogy nem külön szóról van szó, ezt összevonással lehet korrigálni. Természetesen ezzel az előfordulási arányok is változni fognak. Az alábbi ábrán a *jutalom* és a *jut* szavak különböző alakjait láthatjuk.

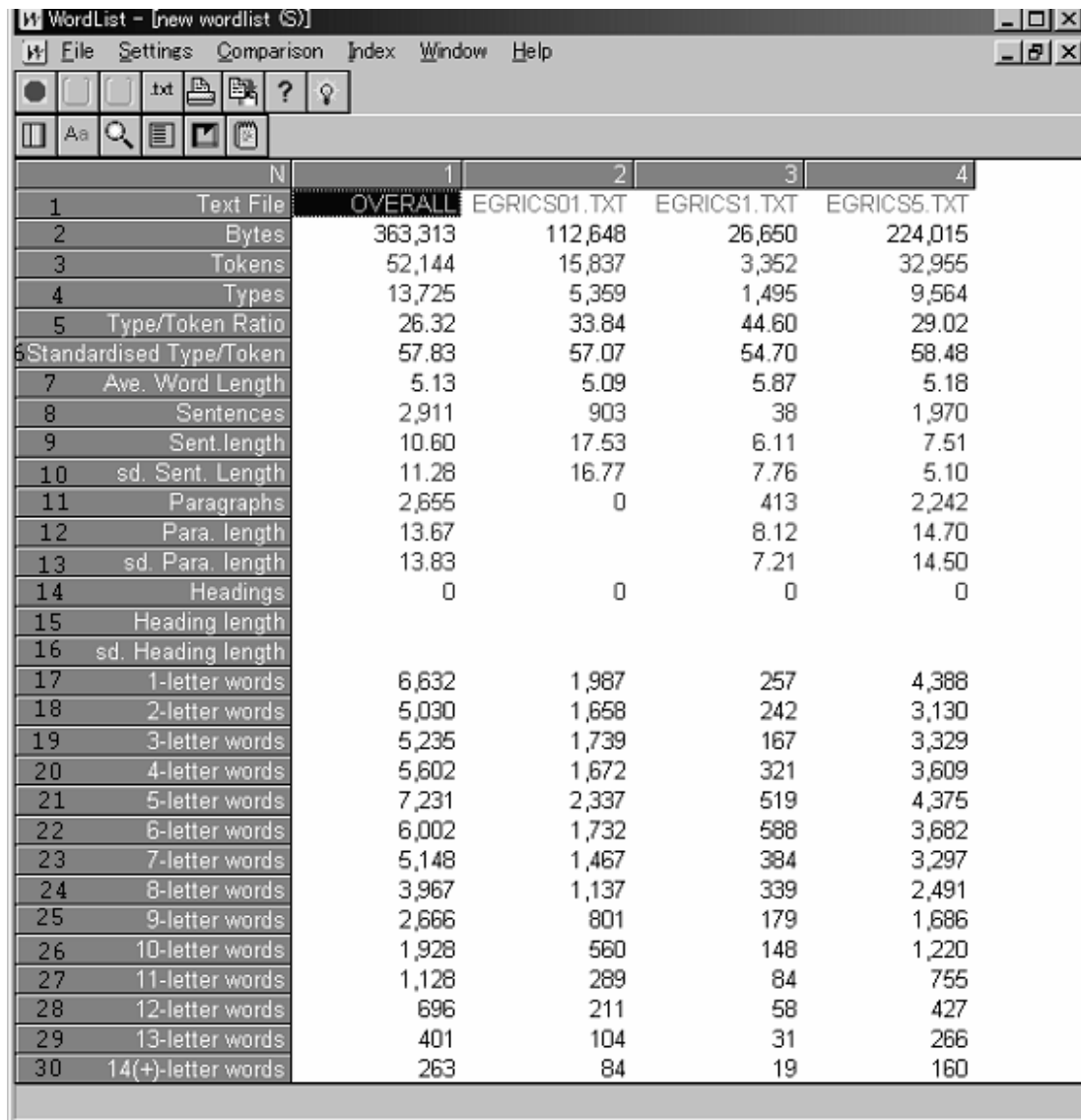
N	Word	Freq.	%	Lemmas
5667	JUTALMAZLAK	1		
5668	JUTALMAZTA	2		
5669	JUTALMUL	1		
5670	JUTALOM	2		
5671	JUTALOMMAL	1		
5672	JUTALOMRA	1		
5673	JUTHATTAM	1		
5674	JUTNAK	2		
5675	JUTNIA	1		
5676	JUTNOM	1		
5677	JUTNUNK	1		
5678	JUTOK	1		
5679	JUTOTT	7	0.01	
5680	JUTOTTAK	6	0.01	
5681	JUTATTANI	2		

32. ábra: Ábécé szerinti szólista (WordSmith program)

A statisztikai adatok programonként változnak. Nézzük meg, egy közkedvelt program, a WordSmith²⁵ (M. Scott, 1996) milyen adatokkal szolgál. Az alábbi ábra mellett található számok az információ azonosítását segítik.

1. a szövegfájl neve
2. mérete bájtokban
3. szövegszavak száma (összes szó a szövegben)
4. különböző szóalakok
5. különböző szavak és összes szó aránya
6. az 5 pont standardizált változata
7. betűk átlagos száma egy-egy szóban
8. mondatok száma
9. a mondatok átlagos szószáma
10. a 9. pont standardizált változata
11. bekezdések száma
12. bekezdés átlagos hossza
13. 12 pont standardizált változata
14. címsorok
15. címsorok átlagos hossza
16. 15 pont standardizált változata
- 17–30-ig az 1, 2, 3 stb. betűből álló szavak száma

²⁵ A WordSmith program nagyon népszerű az egyéni kutatók és tanárok körében, jelenleg kb. 50 angol font az egy számítógépen futtatható változata, melyet a készítője honlapjáról <http://www.lexically.net/wordsmith/> vagy az Oxford University Press lapjáról <http://www.oup.co.uk/isbn/0-19-459400-9> lehet demo változatban letölteni, és a regisztrációs kódot megrendelni. Jelen könyvben kizárólag az ára miatt nem írjuk le részletesen e programot, hanem helyette ingyenesen használhatók kerülnek bemutatásra. A könyv írásakor a program 3. változatát használtuk, a legújabb a 4. változat.



N		1	2	3	4
1	Text File	OVERALL	EGRICS01.TXT	EGRICS1.TXT	EGRICS5.TXT
2	Bytes	363,313	112,648	26,650	224,015
3	Tokens	52,144	15,837	3,352	32,955
4	Types	13,725	5,359	1,495	9,564
5	Type/Token Ratio	26.32	33.84	44.60	29.02
6	Standardised Type/Token	57.83	57.07	54.70	58.48
7	Ave. Word Length	5.13	5.09	5.87	5.18
8	Sentences	2,911	903	38	1,970
9	Sent. length	10.60	17.53	6.11	7.51
10	sd. Sent. Length	11.28	16.77	7.76	5.10
11	Paragraphs	2,655	0	413	2,242
12	Para. length	13.67		8.12	14.70
13	sd. Para. length	13.83		7.21	14.50
14	Headings	0	0	0	0
15	Heading length				
16	sd. Heading length				
17	1-letter words	6,632	1,987	257	4,388
18	2-letter words	5,030	1,658	242	3,130
19	3-letter words	5,235	1,739	167	3,329
20	4-letter words	5,602	1,672	321	3,609
21	5-letter words	7,231	2,337	519	4,375
22	6-letter words	6,002	1,732	588	3,682
23	7-letter words	5,148	1,467	384	3,297
24	8-letter words	3,967	1,137	339	2,491
25	9-letter words	2,666	801	179	1,686
26	10-letter words	1,928	560	148	1,220
27	11-letter words	1,128	289	84	755
28	12-letter words	696	211	58	427
29	13-letter words	401	104	31	266
30	14(+)-letter words	263	84	19	160

33. ábra: Statisztikai adatok az *Egri csillagok* 3 különböző fájljáról

Természetesen arra is kíváncsiak lehetünk, hogy milyen szavak szerepelnek a keresett szóval együtt. Ez nem feltétlenül a közvetlenül mellette levő pozíciót, hanem egy általunk meghatározott „távolságot” jelent. Például a *török** alak keresése esetén a *török*, *törököt*, *töröknek*, és egyéb *török* alakkal kezdődő szó is keresett szóként szerepel. Az általunk meghatározott távolság 5 szónyi a keresett szótól balra és jobbra. Arra vagyunk kíváncsiak, hogy milyen szavak szerepelnek leggyakrabban a *török**-kel együtt. A keresés eredménye táblázat formájában így néz ki:

N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	A	881	595	286	76	65	54	37	363	0	22	61	72	63	68
2	TÖRÖK	426	18	17	7	2	3	1	5	391	0	2	4	6	5
3	AZ	96	44	52	13	10	14	7	0	0	2	12	12	11	15
4	HOGY	86	55	31	5	8	8	31	3	0	0	6	9	12	4
5	ÉS	80	34	46	12	9	6	4	3	0	0	17	8	11	10
6	IS	70	30	40	4	8	9	8	1	0	9	11	5	9	6
7	S	59	29	30	7	6	8	7	1	0	0	12	6	6	6
8	NEM	58	22	36	11	4	5	1	1	0	9	7	5	8	7
9	EGY	57	33	24	3	3	8	10	9	0	3	7	4	5	5
10	TÖRÖKÖT	48	4	0	2	2	0	0	0	44	0	0	0	0	0
11	TÖRÖKÖK	44	2	1	0	0	0	2	0	41	0	0	0	1	0
12	MEG	38	20	18	6	3	5	6	0	0	2	2	5	2	7
13	VOLT	36	17	19	4	4	6	3	0	0	2	4	4	7	2
14	CSAK	33	12	21	3	0	5	2	2	0	1	3	9	3	5
15	DE	31	9	22	1	1	1	5	1	0	0	12	3	4	3
16	MÁR	24	9	15	2	2	2	3	0	0	2	2	3	6	2
17	TÖRÖKNEK	24	4	0	1	2	0	1	0	20	0	0	0	0	0
18	BÁLINT	23	1	22	0	1	0	0	0	0	20	1	1	0	0
19	DOBÓ	22	11	11	4	3	2	2	0	0	0	3	1	4	3
20	OTT	22	10	12	3	3	1	3	0	0	3	3	1	3	2
21	AZT	20	10	10	2	2	4	2	0	0	1	3	0	4	2
22	HÁT	19	7	12	2	2	2	1	0	0	0	1	6	1	4
23	VAN	18	8	10	1	1	3	2	1	0	3	0	3	4	0
24	KI	17	10	7	0	4	2	4	0	0	0	0	3	2	2
25	KIS	17	13	4	0	1	2	1	9	0	0	0	3	1	0
26	MIKOR	17	10	7	2	2	2	4	0	0	0	1	1	3	2

34. ábra: A török* szöveggörnyezetében előforduló szavak (WordSmith)

A táblázat első oszlopában szerepel a szó, a második oszlopban az összes előfordulások száma, a harmadik oszlopban a keresett szótól balra való előfordulások száma, a negyedik oszlopban a jobbra való előfordulásoké. Mivel a jobbra és a balra pozíció egytől öt szó távolságig terjed, fontos tudni, hogy milyen közel vagy távol kerülhet ez a szó a keresett szótól. Az ötödiktől a kilencedik oszlopig a balra elfoglalt hely szerinti szám található, a tizedik oszlop a keresett szót jelzi, a tizenegyedikől a tizenötödikig pedig a jobbra levő hely szerinti előfordulás száma látható. Hozzá kell még tennünk, hogy ezek az adatok a mondatátárokat figyelmen kívül hagyják. Az ilyen jellegű információk azonban nagyon fontosak a kollokációk tanulmányozásában. Meg kell azonban jegyeznünk azt is, hogy a *török* szót nemcsak főnévként és melléknévként, hanem a *tör* ige egyes szám első személyű alakjaként is használhatjuk. A 34. ábra eredményei azt sejtetik, hogy itt nem igei értelemben szerepel a *török*, hiszen 363

alkalommal közvetlenül előtte határozott névelő áll. Más esetekben a pontos elemzés érdekében az adott szövegkörnyezet megtekintésével tudjuk csak eldönteni vagy ellenőrizni, hogy főnévi, melléknévi vagy igei értelemben szerepel-e az adott szó.

A konkordanciaprogramoknál olyan funkciót is használhatunk, amely lehetővé teszi, hogy a több szóból álló, de ismétlődő kifejezésekre keressünk. Például az előbb említett *török** milyen két másik szóval alkot kifejezést? A legszembevetőbb példák ez esetben *a török tábor* és változatai, valamint *a török kezére* és változatai voltak. A kereséskor nem adtuk meg, hogy a *török* toldalékmentes alakja esetében az igei jelentést a program figyelembe vegye-e vagy sem, tehát erre az alakra kerestünk. Az eredmény szempontjából ez azonban nem is lényeges, mivel a számokból egyértelműen kiderül, ebben a szövegben a *török* általában jezőként szerepel.

Vannak szavak, amelyeket csak bizonyos szavakkal együtt használhatunk, de azok szinonimájával már nem. Talán sokan emlékeznek még Brachfeld Siegfried paródiájára, amelyben a *dugóhúzó*ból *rejtővonó* lett, hiszen a művész logikája szerint a *dug* és a *rejt*, meg a *húz* és a *von* szinonimák, így akár fel is cserélhetjük őket. (A pontosság kedvéért jegyezzük meg, hogy valójában úgynevezett közeli szinonimák, amelyeknél a felcserélhetőség nem feltétlen kritérium.) Ha idegen nyelven beszélünk, mi is követhetünk el ehhez hasonló hibákat, ha nem a megfelelő szavakat válogatjuk össze, vagy ha nem a megfelelő sorrendben használjuk őket. A magyar nyelvben *fekete-fehér* televízióról beszélünk, amit sok nyelvben ugyanilyen módon, a feketét előre helyezve fejeznek ki, pl. az angolban: *a black and white TV*, franciául: *une télévision en noir et blanc*, a németben: *das Schwarz-Weiß-Fernsehen*. A sok példa ellenére azonban óvakodnunk kell az általánosításoktól, hiszen a japán nyelvben ezt pont fordítva használják: 白黒テレビ (*shiro kuro terebi*). Ha a példák láttán esetleg valaki arra a következtetésre jutna, hogy a keleti nyelvekben ezt akkor nyilván fordítva mondják, akkor egy kínai példával azonnal óvatosságra intjük. A kínai nyelvben ugyanis, a magyarhoz hasonlóan, a fekete áll elől: 黑白電視 (*heibai dianshi*).

A „kollokáció” néven ismert, a nyelvtanulás és tankönyvírás szempontjából is jelentős fogalom ebben és a következő fejezetben is többször előfordul, így érdemes e fogalmat itt pontosabban meghatározni. Ez talán azért is szükséges, mert fontossága ellenére a magyar nyelvű szakirodalomban alig található meg e kifejezés. A magyar szerzők által készített *Nyelvi fogalmak kyszótárában* (Kugler & Tolcsvai Nagy, 2000) nem találni ilyen szócikket, mint ahogy sem a *Magyar nyelv kézikönyve* (Kiefer, 2003), sem pedig *A nyelv és a nyelvek* (Kenesei, 2004) indexében sem található meg, annak ellenére, hogy a kötetben szerepelnek a szöveggel, gépi szövegfeldolgozással és nyelvtechnológiával foglalkozó írások. Az angol nyelvből fordított *A nyelv enciklopédiájában* (Crystal, 2003: 138) azonban már megtaláljuk, hiszen az eredeti műben is szerepel. Az angol nyelvű szakirodalom bővelkedik kollokációkkal foglalkozó könyvekben és cikkekben, és a kutatások eredményeit egyre több kollokációs szótár készítéséhez használják fel.

Kollokáción bizonyos szavak gyakori együttes előfordulását értjük, de ez nem feltétlen jelenti, hogy a kollokációs (kollokációt képező szavak) közvetlenül egymás mellett állnak, hanem egy bizonyos „távolságon” belül. A szavak természetesen esetlegesen is szerepelhetnek együtt, így joggal merülhet fel a kérdés, hogy hogyan lehet

meghatározni azt a gyakoriságot, amelytől bizonyos szavak együttes előfordulását kollokációnak tekinthetjük. Bonyolult statisztikai képletek és valószínűségszámítási módszerek állnak ehhez rendelkezésre, melyeket szerencsére nem szükséges az olvasónak megtanulni ahhoz, hogy a számítások eredményeit értelmezze. A kollokációk vizsgálatára készült programok már tartalmazzák a számítások elvégzéséhez szükséges kódokat, a felhasználók így azonnal a végeredményt látják.

A kollokáció állandósult kifejezés, de nem idióma, hiszen az idiómák jelentése az alkotóelemek jelentéséből nem áll elő (ez része az idióma definíciójának), pl. a *felkapta a vizet* megértése szempontjából lényegtelen a *felkap* és a *víz* jelentése. Az idiómák általában egy változatban léteznek (nem használjuk azt, hogy **felkapta a tejet* vagy *szörpöt*), így ezeket egy lexikális egységként kezelve könnyen megtanulhatja minden nyelvtanuló. A kollokációkra azonban az jellemző, hogy bizonyos variációs lehetőségek vannak, éppen ezért ezeket sokkal nehezebb a nyelvtanulóknak elsajátítani, mint az idiómákat.

A kollokációk szemléltetésére két melléknevet (*ádáz* és *vad*) vizsgáltunk meg az MNSZ segítségével. A véletlenszerű minta esetében az *ádáz* kollokációiként a következőket találtuk: *csata, ellenállás, ellenfél, ellenség, gyűlölködés, harc, küzdelem*. Jóllehet sok esetben az *ádáz* helyett a *vad* melléknevet is használhatjuk ugyan e főnevekkel (*vad gyűlölködés, ellenség* stb.), de ha a *vad* kollokációit is megvizsgáljuk, akkor azt tapasztaljuk, hogy a fenti kifejezések jelentősen gyakrabban fordulnak elő az *ádázzal*, mint a *vaddal* (pl. *ádáz harc* a teljes korpuszban 64-szer fordult elő, de *vad harc* csak 6-szor). A *vad* nagyon sok különböző szóval szerepelt együtt, kevés volt az ismétlődő még az *ádáznál* 5-ször nagyobb minta esetében is. A többször előfordulók közül említünk meg néhányat: *dolgok, gyönyör, hullámozás, indulat, kíváncsiság, robbanás, rohanás és szenvedély*.

A kollokációk jelentőségét a J. R. Firth vezette londoni iskola ismerte fel elsőként (Firth 1957: 196), és az első kollokációs szótárt is az iskola egyik kiváló képviselője, Harold E. Palmer készítette az 1930-as években (több változatban is). A sok fontos kollokációs szótár közül meg kell említeni Kjellmer (1994) vaskos művét, de Benson & Benson (1993) fontos például az oroszul tanulók számára (és persze az „egzotikus” nyelvek kollokációs szótárait is illene megemlíteni, pl. al-Hafiz 2003-as 373 oldalas arab-angol szótárát – ISBN 9953333793, illetve a kínai Wang Yong és Xie Guofeng 2001-es, 462 oldalas művét – ISBN 7542615084).²⁶

4.3.1. A kezdet kezdetén

A kilencvenes évek elején, amikor még kevesen ismerték és használták a konkordancia-programokat, számos szótárt kiadó cég is készített egyszerű, de nem igazán olcsó konkordancia-programot a kísérletező pedagógusok számára. Abban az időben ez természetesen DOS-ban működő programot jelentett. Nyilvánvaló, hogy a nyelvészek és a nyelvtanárok igényei mások. Így a könnyen kezelhetőség és a világos, könnyen átlátható és szerkeszthető programok lettek népszerűek. A legismertebb programok a következők voltak:

²⁶ Köszönet jár Cseresy Lászlónak az „egzotikus” információért.

- A **Longman Mini-Concordancer** (1989), mely képes volt a szavak számát meghatározni, de viszonylag kis méretű fájlokkal dolgozott (kb. 65 000 szó volt a maximális fájl méret. Talán többen is találkoztak már Chris Tribble és Glyn Jones (1990) könyvével, melynek címe *Concordances in the classroom*, és amely mintegy tanári kézikönyvként szolgált ehhez a programhoz.
- **Micro-OCP** („Micro-OCP”, 1988)
- **WordCruncher** (Brigham Young, 1989)
- **Tact**
- **Clan**
- **Free Text Browser**
- **MicroConcord** (M. Scott & Johns, 1993), melynek minimális igénye az MS-DOS 3.0 változata, kb. 200K RAM és 5,25 vagy 3,5 inches hajlékonylemez meghajtó. A program mindössze 156Kb.

Minden fent említett program Windows alapú. A Macintosh programok között azonban már ekkor is megtalálhatóak voltak az ingyenes programok. Mivel Magyarországon a Windows operációs rendszerek sokkal elterjedtebbek, mint a Macintosh rendszerek, a Macintosh rendszereken futó programokat csak érintőlegesen említjük.

Nem hiszem, hogy sok értelme lenne MS-DOS programok leírásával tölteni az időt, hiszen történelmi jelentőségükön kívül semmi gyakorlati haszon nem származik belőle. Senki nem rohanja meg a boltokat, hogy MS-DOS programot vegyen, még akkor sem, ha valóban jól működtek. Az újabb programok olcsóbbak és tetszetősebb felhasználói felülettel rendelkeznek. Nem beszélve arról, hogy a szoftverek készítésekor az eddigi kutatások eredményeit igyekeznek figyelembe venni, és a lehetőségekhez képest azokat úgy alakítani, hogy azok az új kutatási és számítástechnikai igényeknek megfeleljenek. Az igen kedvelt MS-DOS alapú MicroConcord program Windows XP operációs rendszeren már sajnos nem is működik. Az internet jelentőségének megnövekedésével és a „tömegterjesztés” lehetőségével az egyénileg gyártott ingyenes vagy olcsó programok is fellelhetők az interneten.

4.3.2. Internetes felületen futó ingyenes programok

Számos olyan ingyenes program létezik, amely lehetővé teszi vagy az adott programhoz tartozó korpuszban való keresést, vagy pedig a saját számítógépünkön levő fájlban való keresést a program letöltése nélkül. Jó példa erre a Web Concordancer nevű program (<http://www.edict.com.hk/concordance/default.htm>), mely számos különböző korpuszban való keresés lehetőségét nyújtja. A Bibliától kezdve Drakuláig, a The Times egyes számaitól a „standard” LOB, Brown Korpuszig sok minden megtalálható itt. A keresés eredménye mellett egy szótárhoz vezető kapcsolat is található, amely segít a szavak jelentésének megértésében. Nem véletlen, hogy az angol meghatározás mellett a kínai jelentést is megtaláljuk, hiszen a honlap címéből is kitűnik, hogy Hongkongban került az internetre ez a program. Az alábbi ábra a *house* szó előfordulását mutatja a *The Times* napilap 1995 januárjában megjelent számaiban. A keresett szó 2001-szer szerepel ebben a korpuszban.

Web Concordancer is now searching corpus **TimesJan95.txt** for **house**

Concordances for house = 2001	Net Dictionary entries for house
1 't carry a tune from a well to the house in a bucket" the boys would never	
2 acs, general director of the opera house , said: „A decaying theatre in Cov	
3 Scottish businessman, who let the house as a holiday sporting estate. Mr	
4 d their sleeping bags to the White House . For a man of Robin Renwick's res	
5 's Wells Ballet reopened the Opera House with a performance of The Sleepin	
6 s in the committee corridor of the House have a rough ranking for the Trad	
7 , or a combination of these? Tweed House is a warning to all judges of arc	
8 r with people who have purchased a house with a bit of land and want somet	
9 Northern Electric from Trafalgar House is a challenge to Nigel Lawson's	
10 describes a year in his life as house -husband: a chap who cleans, shops,	
11 as been ploughed over and a nearby house , then a concrete skeleton, has si	
12 After a passionate debate the House approved a constitutional amendment	
13 ge, a fishing net loft, the engine house of a copper mine, a Victorian lau	

35. ábra: Web Concordancer <http://www.edict.com.hk/concordance/default.htm>

A saját szövegeket „feltölthetjük” erre a keresőre, de a magyar nyelvben használt ékezetes betűk miatt nem lesz ideális az eredmény, hiszen ez a program a legtöbb ingyenesen elérhető programhoz hasonlóan, az angolra és esetleg a programozó által beszélt vagy tanult nyelvekre lesz „kihegyezve”. Könnyebb olyan ingyenes programot találni az interneten, amely gond nélkül kezeli a japán, kínai vagy koreai írásjeleket, mint a magyarral megbirkózót.

A Brit Nemzeti Korpuszt is hasonló módon használhatjuk a következő címen: <http://thetis.bl.uk/lookup.html>, természetesen angol szavak kontextusban való megjelentésére. A Collins Wordbanks Online olyan szolgáltatás, amely lehetővé teszi a Collins Word Web-en rendelkezésre álló korpuszok nyelvi adatainak kutatását. A szolgáltatásért fizetni kell, de a Corpus Concordance Sampler egy 56 millió szavas angol nyelvű korpuszban való ingyenes keresést tesz lehetővé. Ennek használatakor csak 40 konkordanciát láthatunk, de ez is elegendő lehet sok esetben a nyelvtanár vagy tanuló számára (<http://www.collins.co.uk/Corpus/CorpusSearch.aspx>).

A két legnagyobb magyar nyelvű korpusz, a Magyar Nemzeti Korpusz és a Magyar Irodalmi és Köznyelv Nagyszótárának Korpusza / Magyar Történeti Korpusz is szabadon kereshető az internetes keresőoldalon, de sem a korpuszt nem lehet letölteni, sem pedig saját dokumentumot a keresőben futtatni.

4.4. Konkordanciák készítése

Ebben a részben azt mutatjuk be lépésről lépésre, hogy hogyan és milyen programokkal lehet egy vagy több rendelkezésre álló magyar vagy más nyelvű szöveget konkordancia-programok segítségével nyelvi szempontból megvizsgálni. Egyre több olyan program készül, amelynek keresési funkciói és szolgáltatásai megközelítik a régebben csak „profi” intézmények által megfizethetőek szintjét, ugyanakkor már magánemberek

számára is elérhetővé váltak. A viszonylag olcsó és népszerű programok közül a következőket emelnénk ki: Wordsmith Tools 3-as és 4-es változatát (M. Scott, 1999, 2004); a Michael Barlow által készített MonoConc, MonoConc Pro és ParaConc programokat (lásd Barlow, 1999); és a Concordancer (Watt, 2004) nevű programot. Mivel az olvasót most arra kérjük, hogy a fejezet további részét olvasva maga is próbálja ki az itt leírtakat, fontosnak tartottuk, hogy a bemutatásra kerülő programok mindegyike ingyenesen letölthető legyen az internetről. Nagy számuk ellenére kevés azonban az olyan ingyenes konkordanciaprogram, amely alkalmas lenne a magyar nyelvű szövegek vizsgálatára.

Hosszas keresgélés után négy olyan programot választottunk, amelyek különböző igényeket elégíthetnek ki attól függően, hogy milyen célra kívánjuk használni vizsgálatunk eredményeit. Máris felhívnánk a figyelmet arra, hogy a pedagógiai alkalmazásokról a következő fejezetben lesz szó, itt csak esetlegesen és röviden utalunk ezekre. A négy program a következő: 1. ConcApp (Greaves, 2003); 2. Simple Concordance Program (SPC) (Reed, 2003); 3. AntConc (Anthony, 2004); és 4. Multi-Lingual Corpus Toolkit (MLCT) (Piao, 2002). Mindegyik programot magyar Windows XP operációs rendszeren futtattuk, és probléma nélkül működtek. Eredetileg azonban nem feltétlenül erre a platformra készültek, de XP környezetben is működnek. Néhány esetben a korábbi változat(ok), mint pl. Win 98 vagy Win 2000-re írottak most is letölthetők. A következő táblázat a legfontosabb, letöltéshez szükséges információkat tartalmazza. Ha elegendő helyel rendelkezünk a számítógépen, érdemes mindegyiket letölteni és kipróbálni. A <http://lingo.lancs.ac.uk/devotedto/corpora/software.htm> honlapról mindegyik elérhető a Free Concordancer címszó alatt.

Program neve	SCP	AntConc	ConcApp	MLCT
zip fájl mérete	10,8Mb	2,69Mb	teljes 2,55Mb	474kb
utolsó változat	2003	2004	2003	2002
programozó	Alan Reed	Laurence Anthony	Chris Greaves	Scott Songlin Piao
honlap	http://www.textworld.com/	http://www.antlab.sci.waseda.ac.jp/	http://www.edict.com.hk/pub/concapp/	http://www.lancs.ac.uk/staff/piaosl/research/download/download.htm
e-mail cím	A.Reed@textworld.com OR A.Reed@talk21.com	anthony@waseda.jp	ecchgr@hotmail.com	s.piao@lancaster.ac.uk

24. táblázat: Ingyenesen letölthető programok

Minden zip kiterjesztésű fájlt a *winzip* nevű program, vagy más zip kiterjesztésű fájl kezelésére alkalmas programmal kicsomagolunk egy általunk választott nevű könyvtárba. Az SCP program azonnal telepíthető változatban is letölthető. Meglehetősen nagy fájl, így letöltése hosszabb időt vesz igénybe még kábeles internetes kapcsolat esetén is, kb. 20 percbe telt letöltése ebben az esetben. Az AntConc programot letöltése után

azonnal lehet használni, semmilyen telepítésre nincs szükség. A ConcApp programot a setup programmal installáljuk, ezután már a szokásos módon a Start, majd Programok menüpontra kattintva választhatjuk ki és indíthatjuk.

Az MLCT program esetében a program működéséhez szükséges, hogy számítógépünkön legyen egy Java Runtime Environment (JRE) nevű program, amelynek letöltéséhez egy linket találunk a honlapon. A zip fájl kibontása után a mappában levő run_mlct_concordance_jar.bat fájlra való kattintással lehet a felhasználói felületet elindítani. Az első kinyíló ablak azonban egy fekete DOS ablak, és csak egy kis idő elteltével fog megjelenni a második, a tényleges program ablak.

A programok elindítását megkönnyíti, ha a telepítés során létrehozott ikonokat keresünk, mert ezekkel lehet a programokat elindítani. A konkordanciaprogramokra általában jellemző, hogy .txt kiterjesztésű fájlokkal működnek. Az újabbak XML, vagy a program súgójában ismertetett egyéb fájl formátummal is működhetnek. A programok tanulmányozása során a Magyar Elektronikus Könyvtárból letöltött *Egri csillagok* szövegfájljaival dolgoztunk. Ajánlott a kisebb méretű fájlokon való kipróbálás, hiszen így gyorsabban meggyőződhetünk arról, hogy egy-egy utasítás kiadása után milyen eredményt kapunk. Ezért sokszor az öt fájlból álló teljes szöveg helyett csak egy szövegfájlt használtunk.

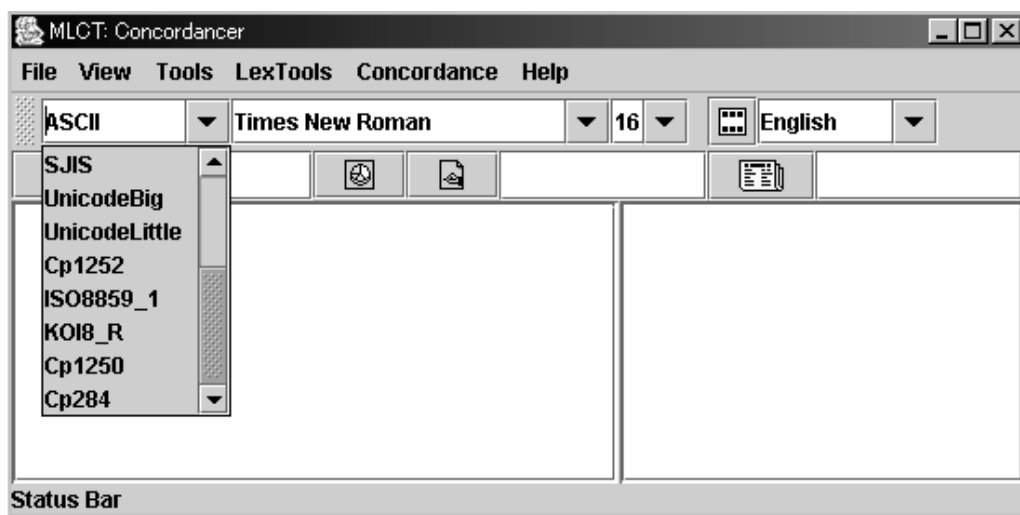
Az itt felsorolt programok mindegyike folyamatos fejlesztés alatt áll, így előfordulhat, hogy ha ma még nem is tudunk valamilyen elemzést elvégezni a programmal, a következő hónapban viszont már igen. Ezért javasoljuk, hogy az olvasó időnként „nézzen vissza” a program honlapjára és töltsen le a legújabb változatot. Még egy általánosnak nevezhető dologra kell felhívni a figyelmet. A konkordanciaprogramok készítése során egyre pontosabb keresési módokat építenek a programokba készítőik. Itt a regular expression, röviden regex vagy regexp használatára gondolunk. A keresés módjánál sokszor választható, hogy szavakat keresünk-e vagy regexeket. Egy szót is be lehet gépelni regexként. Ezt használtam ki, amikor egy program nem tudta értelmezni a magyar ékezetes betűket. Ha szavakat kerestem, az ékezetes betűket teljesen figyelmen kívül hagyta a program a keresésénél, de ha regexként írtam be az ékezetes szót, akkor „megtalálta”. Elégedjünk itt meg azzal a meghatározással, hogy a regex olyan szimbólumok és szintaktikai elemek készlete, amelyekkel szövegszerkezeteket (pattern) azonosíthatunk. Példaként említsük a két leggyakrabban használtat: bizonyára sokan használták már kereséskor a ?²⁷ vagy a *²⁸ szimbólumot, de említhetnénk a keres és cserél funkciót is az MS Word használatakor. Nyilvánvaló, hogy ezek segítségével pontosabb kereséseket végezhetünk a szövegben. Itt nem foglalkozunk ismertetésükkel, de érdekesnek tartjuk egyéni tanulmányozásukat.

²⁷ A ? egy karakter, azaz betű vagy szám helyettesítésére alkalmas jel. Például, ha a *t?r* kifejezésre keresünk, a ? helyén állhat bármilyen betű, így a *tar, tár, tér, tör, tór, túr, tūr* alakok mindegyikét egyszerűen megtaláljuk.

²⁸ A * tetszőleges számú karaktert helyettesít. Például a *török** keresés eredményeként megtaláljuk a *török, törökül, törökök, töröknek* stb. alakokat.

4.4.1. Az MLCT

Kis mérete ellenére talán a legmélyrehatóbb lexikai vizsgálatokat az MLCT programmal végezhetjük. Ez a program része egy programkészletnek, amely többnyelvű szövegfeldolgozásra és nyelvi vizsgálatok céljára készült. A program indítása után két dologra kell figyelni. A program által használt kódolást, a beállított ASCII-ról Cp1250 vagy Cp1252-re át kell állítani. Ha ezt elmulasztjuk még a fájl megnyitása előtt, az ékezetes betűk nem fognak helyesen megjelenni a képernyőn. A program kétnyelvű dokumentumok vizsgálatára is különösen alkalmas, hiszen két ablakkal működik, és kívánság szerint a jobb vagy a bal oldalon nyithatjuk meg a dokumentumokat. Ha egy nyelvvel vagy fájlal dolgozunk, akkor a bal oldalon nyissuk meg a fájlt, mert az eredményeket mutató ablak a jobb oldali. Az alábbi ábra mutatja a kezdő ablakot és az ASCII megváltoztatására szolgáló menüt. A kódolásokból látható, hogy japán, koreai és kínai szövegeket is vizsgálhatunk a programmal, ha a számítógépünkre a megfelelő kiegészítő programok telepítve vannak, amelyek az ezeken a nyelveken történő szövegszerkesztéshez is elengedhetetlenek. Az ábrán a Times New Roman betűtípus neve olvasható. A jobb oldalon levő nyílra kattintva a legördülő lehetőségek közül kiválaszthatjuk az általunk kedvelt és a vizsgált nyelvnek legjobban megfelelő betűtípust. A betűk mérete (az ábrán 16 pont) hasonló módon állítható a következő legördülő menü segítségével.



36. ábra: Az MLCT program kezdő ablaka

A fenti ábrán látható, hogy a nyelv jelenleg angolra van állítva, így az angol nyelv szerinti ábécésorrendet és betűkészletet használja a program. A választási lehetőségek: angol, kínai, koreai, finn és egyéb nyelvek. Az ékezetes betűk miatt soha nem fogunk a programtól hibátlan magyar ábécé szerinti listát kapni, de ezt más programban könnyen korrigálhatjuk. A program a kis- és nagybetűket megkülönbözteti. Az English felirattól balra eső, a dobókocka hatos számát mutató ikonra való kattintással a bal ablakban levő

szöveget automatikusan mondatokra és paragrafusokra oszthatjuk, aminek eredményét a jobb oldali ablakban láthatjuk majd. Az ablak alsó részén, ahol a fenti ábrán jelenleg a Status Bar olvasható, a program a műveletek végzése közben az adott műveletet kiírja, majd a művelet végrehajtása után ismét a Status Bar jelenik meg. Egyedül ez alapján tudjuk csak megállapítani, hogy a program még dolgozik-e az adott műveleten vagy már befejezte azt.

A nyitó ablak áttekintése után nézzük meg a menükben szereplő utasításokat. A File (fájl) menü a következőkből áll:

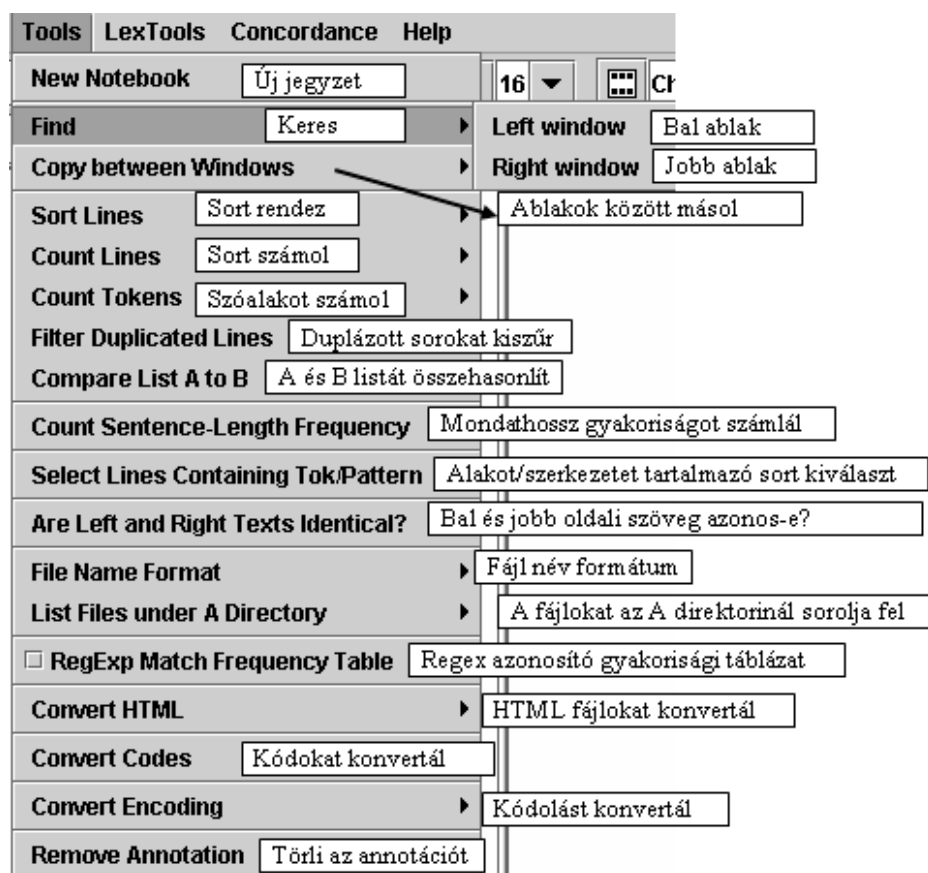
New left window	Új bal ablak
Open in left window	Bal ablakban kinyit
Save left window as	Bal ablak mentése másként
New right window	Új jobb ablak
Open in right window	Jobb ablakban kinyit
Save right window as	Jobb ablak mentése másként
Extract text from webpage	Honlapról szöveget kivon
Exit	Kilépés

37. ábra: Az *MLCT* File menüje

Első lépésként nyilván valamilyen szöveget kell megnyitnunk, tehát a Bal ablakban kinyit menüre kattintva a megfelelő fájlt kiválasztjuk. A fájl formátum választási lehetőségei a következők: egyszerű szövegfájl (txt), Latex dokumentum, HTML, XML és SGML dokumentum. A másik lehetőség a honlapról való szövegkinyerés. Ebben az esetben figyelni kell arra, hogy a számítógépre telepített tűzfal (firewall) a program internetre való kapcsolódását engedélyezze.

A View menüben a jobb és a bal ablakban a szövegnek az ablakhoz való méretezését állíthatjuk, valamint a háttér és előtér színeit. A két következő menü tartalmazza az igazi nyelvi elemzés utasításainak nagy részét, melyek közül több almenüt is tartalmaz. A Tools menü számos almenüje a művelet elvégzési helyének választási lehetőségét kínálja fel, így a jobb vagy bal oldali ablakot. Esetenként a művelet elvégzését, pl. a Duplázott sorokat kiszűr esetében, a fájlokban is felkínálja. A HTML fájlokat vagy egyszerű szövegfájlra vagy a mondat/bekezdés határokat bejelölt szöveggé alakítja. A Conevrt Encoding (átkódolás)²⁹ esetében csak bizonyos kombinációk választhatók a listából, pl. UTF 16-ról UTF 8-ra stb.

²⁹ A számítástechnikában a különböző nyelvekhez tartozó kódlapok feladata az, hogy az adott nyelv karaktereit megfelelően kódolják és jelenítsék meg. Ha például az internetes böngészés során olvashatatlannak jelenik meg egy szöveg, a kódolás állításával, azaz a helyes kódlap kiválasztásával korrigálhatjuk a hibát. Jelen esetben magát a kódlapot változtathatjuk meg.



38. ábra: A Tools menü pontjai

A fenti ábrán a menüpontok önmagukért beszélnek, így külön magyarázatot nem fűzünk ezekhez.

A LexTools menü neve is elárulja, hogy itt komolyabb lexikai jellegű eredményekre számíthatunk. Az első két menüpont a leghasznosabb az átlagos felhasználó számára. Az első, a Remove Punctuation Marks? (Eltávolítja az írásjeleket?), bekapcsolásával vagy kikapcsolásával meghagyhatjuk vagy eltávolíthatjuk az írásjeleket a szöveg vizsgálatakor. Az Extract N-grams (n-gramok kinyerése) pont almenüjéből 1-től 6-ig választhatunk. De mi is az n-gram (ejtsd: engrem)? Ha az 1n-gramet választjuk, akkor egy szólistát kapunk, ahol minden sorban egy szó szerepel. A 2n-gram esetében minden sorban két szó szerepel. Ezek a szavak a szövegben egymás mellett levő szavak. Nézzük meg ezt számokkal szemlélítve. Ha a szöveg szavait 1-től 10-ig terjedő számokkal helyettesítjük, akkor a 2n-gram a következőképpen néz ki:

12 23 34 45 56 67 78 89 910

39. ábra: 2n-gram számokkal szemlélítve

Ha 3n-gramet akarunk vizsgálni, akkor ugyanez a példa a következőre változik:

123 234 345 456 567 678 789 8910

40. ábra: 3n-gram számokkal szemléltetve

Mondathatárokat nem vesz figyelembe az ilyen elemzés, mely arra alkalmas, hogy olyan ismétlődő szerkezetekre hívja fel a figyelmet, amelyeket különben nem vennénk észre. Ennek kipróbálásához igen rövid fájl használatát javasoljuk, mert a program futtatása sok időbe telhet, és ezalatt azt is nehéz megállapítani, hogy a gép működik-e, egyáltalán érdemes-e várni. Az Extended Porter's Stemmer (Bővített Porter szótövező) csak az angol nyelv vizsgálatakor használható, hiszen az angol nyelv sajátosságainak megfelelően szótövekre „vágja” a szöveget. A következő két menüpont a kollokációk vizsgálatánál használható.

A Collocation Parameters (Kollokációk paraméterei) alpontjai a következők:

Update Scan Distance	Keresési távolság frissítése
<input type="checkbox"/> Limit Number of Tokens?	Limitálja a szóalakok számát?
Update Max Number of Tokens	Frissíti a max. szóalak számot
<input checked="" type="checkbox"/> Filter by T-score (1.65)?	T-score (1,65) alapján szűrjön?
<input checked="" type="checkbox"/> Filter by Frequency?	Gyakoriság alapján szűrjön?
Update Min Frequency	Frissíti a min. gyakoriságot

41. ábra: A Collocation Parameters almenüje

Ezek közül a T-score (ejtsd: tíszkór) magyarázatra szorul. Ez olyan statisztikai adat, amely megmutatja, hogy hogyan viszonyul egy-egy kollokáció tényleges előfordulása a valószínű előforduláshoz. Minél nagyobb ez a szám, annál biztosabb a kollokáció előfordulása. Az utolsó menüpont megértéséhez komoly ismeretek szükségesek, ezek meghaladják az átlag felhasználó szükségleteit.

Mutual Information (MI)
Mutual Information (Squared)
Mutual Information (Cubic)
Phi-Square
Log-likelihood
Ochiai
McConnoughy Coefficient
Yule Coefficient
Fager/McGowan Coefficient
Kulczinsky Coefficient

42. ábra: Az Extract Collocates By statisztikai együttthatókat felkínáló alpontjai

A fenti listából csak egy elem emelnénk ki, a Mutual Information-t, amelynél a vizsgált elem és kollokánsa arról adnak kölcsönösen információt, hogy tényleges együttes előfordulásuk hogyan viszonyul a várható előforduláshoz, feltételezve azt, hogy előfordulásuk esetleges. (Lásd McEnery & Wilson, 1996; Oakes, 1998; Ooi Vincent, 1998). Ha egy szó kollokációit különböző statisztikai számítások alapján készítjük, a listákon szereplő szavak vagy azok sorrendje más és más lesz. Ha a fenti statisztikai módszerek alaposabb megismerésére törekednek, Oakes könyvének 4. fejezetét (1998: 149–197) ajánlom további tanulmányozásra. Az angolul nem tudók a 171. oldalon találhatják meg az Ochiai-, McConnougy-, Yule-, Fager/McGowan- és Kulczinsky-féle számítások matematikai képletét.

A Concordances (Konkordanciák) menüben frissíthetjük a paramétereket (szövegekörnyezet hosszát az első pontból), a másodikból a konkordanciák készítéséhez használni kívánt fájlokat választhatjuk ki, a harmadikkal a kiválasztást törölhetjük, és az utolsó pontnál a konkordanciák ábécérendbe való állításának módját választhatjuk ki: balra az első, második és harmadik szó, valamint jobbra az első, második vagy harmadik szó.

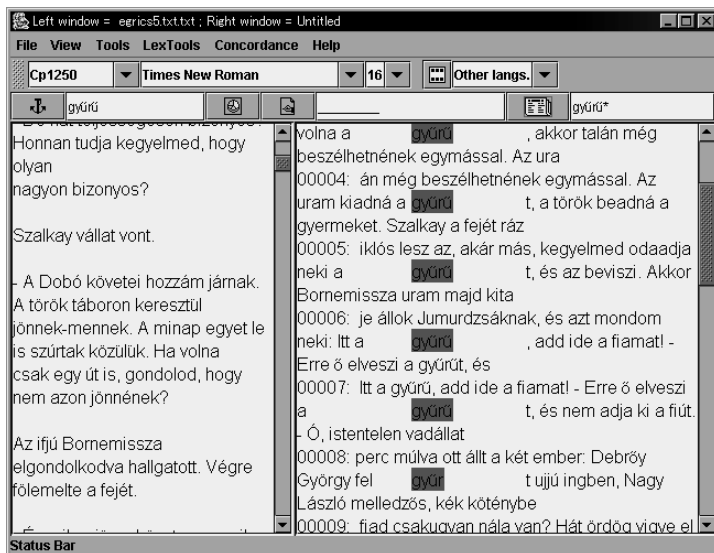
Még négy ikon és három szövegablak szorul magyarázatra, melyeket a következő ábrán láthatunk:



43. ábra: Kontroll ikonok és szövegablakok

Az első szövegablakba beírt kifejezést vagy szöveget a horgony ikonra való kattintással regexként keresi a bal ablakbeli szövegben. Az eredményt a jobb oldali ablakban láthatjuk. Amennyiben a Tools menü RegExp Match Frequency Table (Regex azonosító gyakorisági táblázat) almenüjét bejelöljük, az eredményt rendezhető táblázat formájában is megtekinthetjük.

A kört ábrázoló ikon segítségével az első szövegablakba beírt kifejezést vagy szót a második ablakba beírt kifejezéssel „lecserélhetjük”, azaz helyettesíthetjük. Ha a második szövegablak üres, akkor az első szövegablakban szereplő kifejezést üressel helyettesíti vagy törli. Az eredményt a jobb oldali ablakban láthatjuk.



44. ábra: A programablak konkordanciákkal a jobb oldalon

A kéz ikon arra szolgál, hogy a második szövegablakban megadott módon megváltoztassa a keresett elemet, majd a megváltoztatott keresés eredményét kilistázza a jobb oldalon.

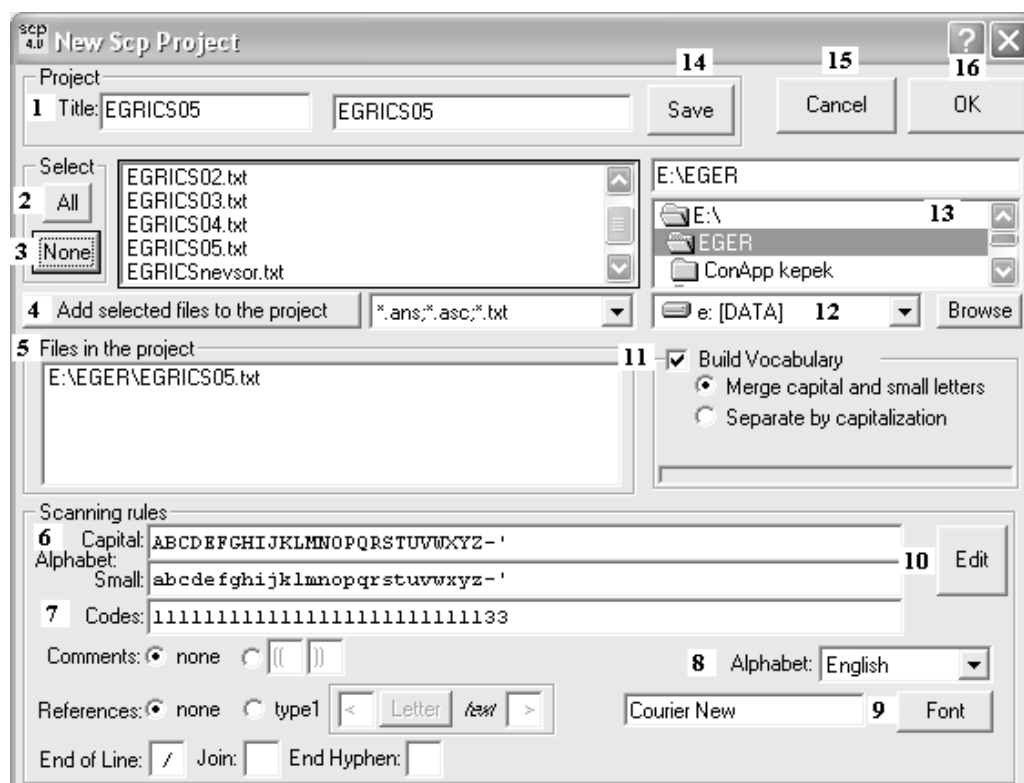
Az utolsó, könyv ikon a harmadik szövegablakba gépelt szöveg vagy regex konkordanciáit keresi ki a bal ablakban levő szövegből és a jobb oldali ablakban teszi láthatóvá. A harmadik szövegablakba gépelt kifejezés zöld színnel lesz feltüntetve a konkordanciákban. A 16. ábra jobb oldali ablakában láthatjuk a *gyűrű* szó konkordanciáját.

4.4.2. Simple Concordance Program SCP

Azért választottuk másodikként bemutatásra ezt a viszonylag nagyobb méretű programot, mert egy kis igazítással a magyar nyelvre „hangolható”, és majdnem a magyar ábécé szerinti listát leszünk képesek létrehozni segítségével. A problémát csak a két betűből álló kapcsolatok okozzák, így tehát a *cs*-vel kezdődő szavak a *cu*-val kezdődők előtt fognak szerepelni, és nem pedig utánuk. Az ékezetes betűk ebben a programban nem okoznak problémát, ha a projekt megkezdésekor a karakterkészletet kiegészítjük a magyar ékezetes betűkkel. Így tehát a magyar szavakat szóként fogja felismerni a program, míg más programok az ékezetes betűket a szavak keresésekor szóközként értelmezik.

A program indítása után megjelenő ablakban első lépésként a File menüből az Open, azaz Megnyit almenüt választjuk. Automatikusan a program mappája (scp32v407) nyílik ki és a program részeként letöltött, már kész scp kiterjesztésű projektek közül választási lehetőséget kínál fel. Ezek angol nyelvűek (2cities, gawain, lincoln), de kísérletezésre és tanulásra talán az angolul nem tudóknak is hasznosak lehetnek. Ha azonnal saját szöveggel akarunk dolgozni, akkor két lehetőségünk van. Vagy ugyanebből az ablakból folytatva a szokásos módon megkeressük a kívánt szövegfájlokat tartalmazó mappát, vagy pedig az Open menü helyett a New választásával azonnal egy új projektet kezdhetünk. Ha az előbbi megoldást választjuk, azaz a felkínált projektek helyett választunk saját fájlt, akkor ne felejtsük el a fájltypust átállítani szövegre, mert különben nem jelennek meg a listán a szövegfájlok. Ekkor ugyan még csak egy szövegfájl választhatunk ki és nyithatunk meg, de a következő ablak kinyílása után a mappában levő összes szövegfájl is kiválaszthatjuk a projekthez. Így javasoljuk, hogy vagy az összes szövegfájl tegyék egy mappába, vagy legalábbis a projekthez használni kívántakat, még a program elindítása előtt. A kétféle kezdési mód ugyanahhoz az ablakhoz vezet, amit a 45. ábra mutat be. Az 1-es számmal jelölt szövegdozba írhatjuk be a projekt nevét. Az ez alatt levő szövegdozban látható az összes szövegfájl, ami a megnyitott mappában szerepel. Ha mindet ki akarjuk választani, akkor csak a 2-sel jelzett All gombra kell kattintani, és automatikusan az 5-ös számmal jelzett dozba ugranak, amely a projektbe felvett fájlokat mutatja. Ha tévedtünk, akkor vagy az összes fájl törölhetjük a projektből a 3-as gomb None megnyomásával, vagy az alsó dozban levő nem kívánt fájlt, a nevére kétszer kattintva. Ha egyesével akarjuk a mappából kiválasztani a projektbe kerülő fájlokat, akkor ebben az esetben is duplán kattintunk a felső dozban a nevére, vagy egy kattintással kijelöljük, és a 4-es gombot választjuk.

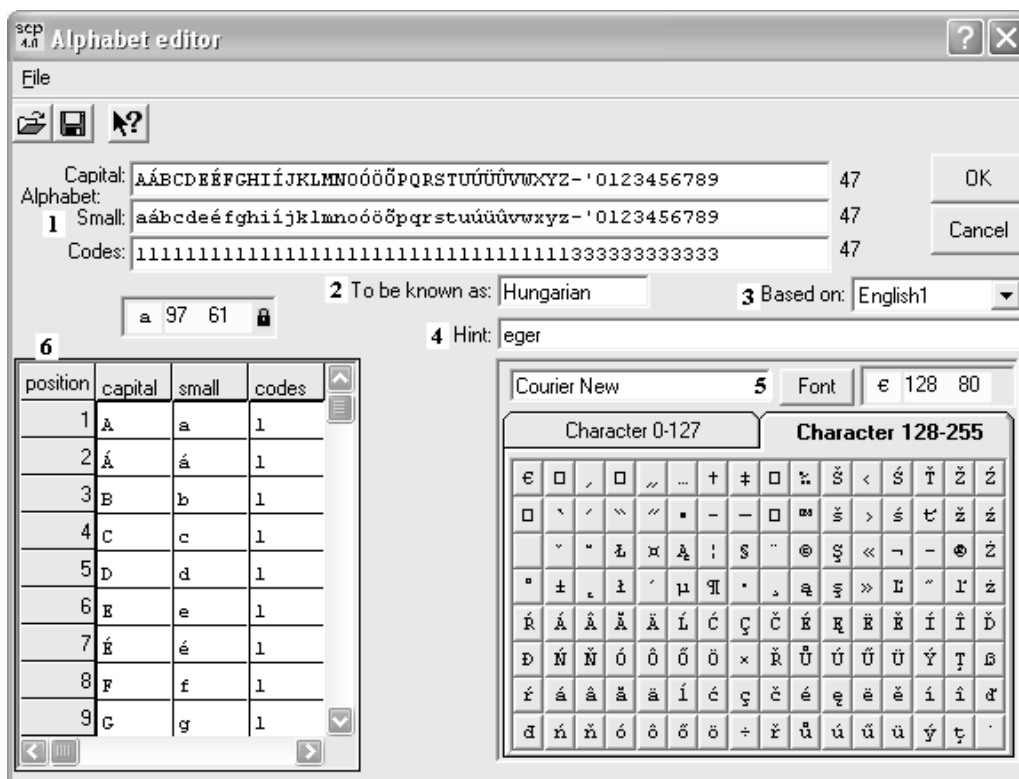
A szövegfájlok kiválasztása után a következő fontos lépés a projekthez szükséges karakterek és betűtípusok megadása. A 8-as számnál láthatjuk, hogy jelenleg angol nyelvre van állítva és a 6-os számnál két sorban szerepel az összes ebben a beállításban használt karakter, felül a nagybetűk, alul pedig a kisbetűk. A 7-es számnál kódokat látunk, melyek a számítógép számára adnak utasítást, hogy hogyan kezelje ezeket a karaktereket. Ha nem magyar, hanem más nyelvet használunk, például orosz, német, franciát, spanyolt, dán, görögöt, izlandit, svédet, arabot, norvéget, hébert vagy katalánt, akkor csak a megfelelő nyelvet kell kiválasztanunk. Természetesen a számítógépünkön meg kell, hogy legyen a megfelelő nyelvi támogatás is. A Font (9-es gomb) megnyomásával egy újabb ablak nyílik meg, ahol kiválaszthatjuk a betűk típusát, stílusát és méretét a megfelelő írásrendszerrel együtt.



45. ábra: Új projekt létrehozása

Mivel a magyar nyelv nem szerepel a választható nyelvek között, elkerülhetetlen a betűkészlet saját kezű szerkesztése. A 10-es gombot megnyomva a 46. ábrán látható ablak nyílik meg. A szerkesztés legegyszerűbb módja, ha egy már létező, a magyar ábécéhez leginkább közel álló karakterkészletet egészítünk ki. Ebben az esetben az English 1-re esett a választás, mely a 3-as gombnál választható ki. Ezek után az 1-gyel jelzett szövegdobozban a megfelelő helyre kattintunk, ahol megjelenik a kurzor. A jobb alsó sarokban levő karaktertáblából az egérrel kiválasztjuk a megfelelő betűt, amely az egyessel jelzett szövegdobozban a kurzor helyén azonnal megjelenik. A nagy és kisbe-

tüket egyesével kiválasztva, ne felejtjük el a kódokat sem beírni! Minden beírt betű kódja 1 lesz. A sorok végén levő számoknak egyezniük kell. A 6-os számmal jelzett oszlopok a fentebb levő szövegdobozzal egyező információt mutatnak, de talán könnyebb itt észrevenni a hibát, mint a felsorolásban. A 2-vel jelzett szövegdobozba írjuk az általunk választott nevet a karakterkészlet számára. A 3-at üresen is hagyhatjuk, vagy valami utalást írhatunk bele. Végül ne felejtjük el az OK gombot megnyomni. Ha valamit elvettünk, a program egy ablak megjelenésével ezt jelzi. Ha nincs probléma, akkor a 45. ábrán látható ablakhoz jutunk vissza.

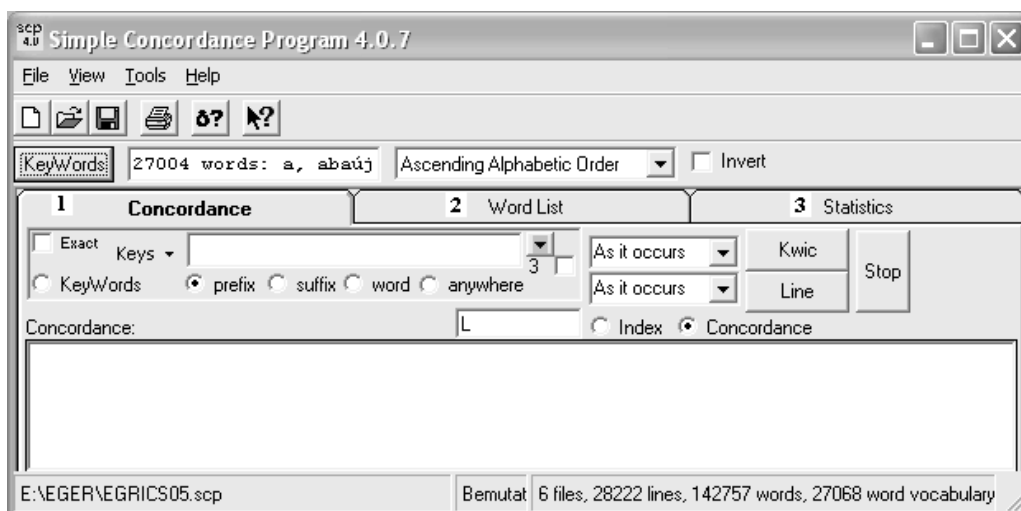


46. ábra: A karakterkészlet szerkesztőablaka

Ha a szövegfájlban nincs magán a szövegen kívül semmilyen megjegyzés vagy referencia, akkor a kódok alatt szereplő választási lehetőségeket az eredeti állapotban kell hagyni.

A 11-es gombnál a Build Vocabulary (Szószedet létrehozása) bejelölésével a program a projekt szószedét automatikusan elkészíti és tárolja. A közvetlenül alatta levő gomb választásával a nagybetűs és kisbetűs szavakat nem sorolja fel külön, hanem azonosnak tekinti, tehát a *kovács* mesterség és a *Kovács* tulajdonnév között nem tesz különbséget. Ha az alatta levőt választjuk, akkor szétválasztja a kis- és nagybetűs szavakat. Ha mindent megfelelően kiválasztottunk, akkor a 12-es számnál kiválasztjuk, hogy melyik merevlemezen, a 13-asnál pedig, hogy melyik mappába akarjuk a projektet elmenteni. Az OK gombra kattintva megjelenik egy ablak, amely azt a kérdést teszi fel, hogy menteni kívánjuk-e a projektet. Mindenképpen ajánlatos elmenteni, legalábbis addig,

amíg jobban meg nem ismerjük a program működését. Az igenre kattintva újabb ablak nyílik, ahol menthetjük a projektet.



47. ábra: A projekt kezdő ablaka

A projekt mentése utáni kezdőablakot a 47. ábra szemlélteti. Három nagy egységet láthatunk: 1. Konkordancia; 2. Szólista; és 3. Statisztika felirattal. A fenti ábra a Konkordancia választási lehetőségeit nyújtja. Mielőtt azonban rátérnénk ennek tárgyalására, figyeljük meg, hogy az ablak alján már most látható néhány fontos statisztikai adat. A projekthez vezető elérési út mellett a projektben szereplő fájlok, sorok, szövegszó és szóalakok száma látható. A Keys melletti szövegdobozba írjuk be a keresett szót, és az alatta levő sorban válasszuk a word kifejezést, majd kattintsunk a Kwic gombra. Ekkor a projekt szövegéből a beírt kulcsszó Kwic formátumban levő előfordulásai az alsó nagy szövegdobozban jelennek meg. Így ha ide a *török* szót írjuk be, akkor csak azokat az előfordulásokat választja ki a program, amelyben a *török* alak áll, de a todalékos szavakat, mint a *töröknek*, már nem. Ha ezekre is kíváncsiak vagyunk, akkor a *törököt*, mint prefixumot kell keresni, annak ellenére, hogy a magyar nyelvben ez nem prefixum. (Erre azért van szükség, mert a program nem a magyar nyelv sajátosságait figyelembe véve készült, hanem az angol nyelv logikáját követi.) Még egy választási lehetőség van. Az anywhere, azaz bárhol választásával is erre az eredményre juthatunk e szó esetében. Azonban ez nem mindig célravezető, például ha az *ér* szót bármilyen előfordulásban keressük, nagyon valószínű, hogy sok olyan szó jelenik meg, amelyben az *ér* olyan módon szerepel, mint például a *denevér* szóban. A konkordanciák megjelenéséig akár egy teljes perc is eltelhet, a fájl méretétől és a számítógép kapacitásától függően, ezért először érdemesebb kisebb fájlon kísérletezni. Ha a Kwic helyett a Line gombot választjuk, akkor a keresett szó nem a sor közepén fog megjelenni, hanem a sor bármely részén, hiszen itt azt a teljes sort láthatjuk, amelyben a keresett szó szerepelt.

A 2-vel jelzett Word List esetében nem sok teendőnk van. A szólista négyféleképpen jeleníthető meg: balra vagy jobbra igazodó oszlopokban, sűrítve, és egy oszlopban. A Layout címszó alatt ilyen sorrendben találjuk meg ezeket. A könnyű áttekinthetőség

érdekében javasoljuk az egy oszlopot. Ezek után már csak a Word List feliratú gombot kell megnyomni, és a szavak ábécé sorrendben, előfordulási számukkal együtt megjelennek. Ha más sorrendben szeretnénk látni az adatokat, például először a leggyakrabban előforduló szót, és csökkenő sorrendben a többit, akkor ezt a Word List felirat feletti legördülő ablakból választhatjuk ki.

A Statistics címszó alatt elég, ha csak egyet kattintunk a Statistics gombra, és automatikusan a 48. ábrán mutatotthoz hasonló információt kapunk.

Word Frequency Profile of 27027 words

Word Frequency	Number of Words	Cumulative Vocabulary	Cumulative Word Count	Percentage Vocabulary	Percentage Word Count
1	17381	17381	17381	64,30976	12,17523
2	3848	21229	25077	78,54738	17,56621
3	1661	22890	30060	84,69308	21,05676
4	925	23815	33760	88,11559	23,64858
5	566	24381	36590	90,20979	25,63097
6	413	24794	39068	91,73789	27,36678
7	298	25092	41154	92,84049	28,82801
8	224	25316	42946	93,66929	30,08329
9	206	25522	44800	94,43149	31,382
10	145	25667	46250	94,96799	32,39771
11	131	25798	47691	95,4527	33,40712
12	97	25895	48855	95,8116	34,22249
13	83	25978	49934	96,1187	34,97837

D:\Program Files\scp32v407\eger.scp EGER 6 files, 28222 lines, 142757 words, 27027 word vocabulary

48. ábra: Statisztikai adatok

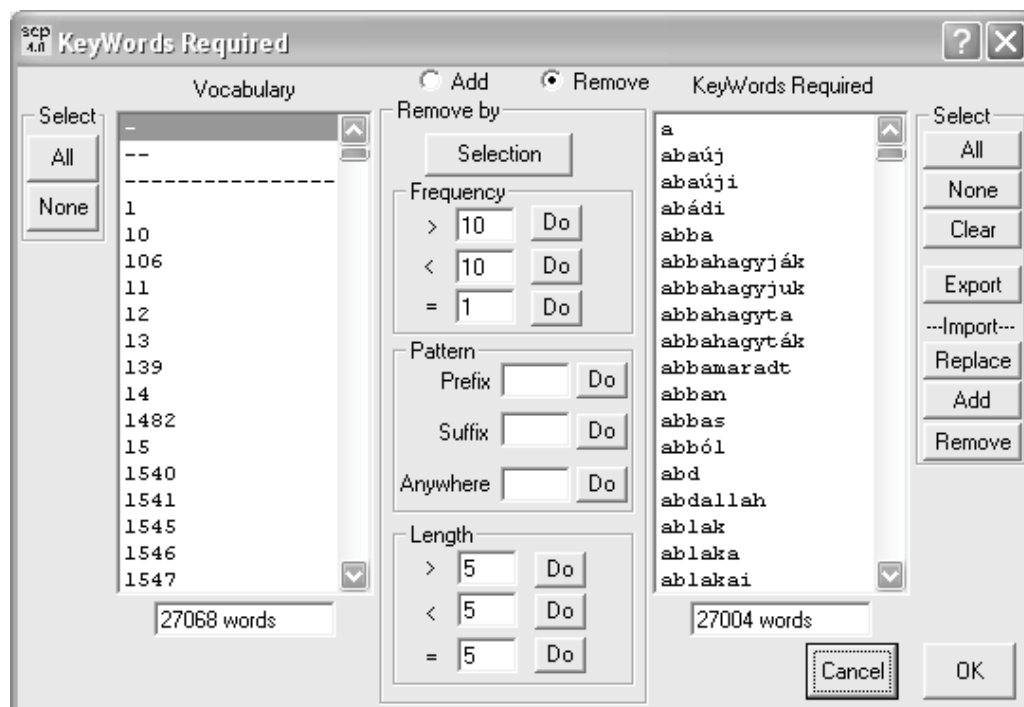
Az első oszlop a szavak gyakoriságát mutatja, tehát az első sorban az egyszer előforduló szavak szerepelnek. A második oszlop az egyszer előforduló szavak számát mutatja. Ebben a projektben 17 381 szó szerepel csak egyszer a szövegben. A harmadik oszlopban azt láthatjuk, hogy ez az eddigiekkel együtt összesen hány szót tesz ki. A harmadik oszlop második sora a második oszlop első és második sorának összege, tehát az egyszer és kétszer előforduló szavak szóalakjának számát mutatja. A negyedik oszlop második sora azt mutatja, hogy az egyszer és kétszer előforduló szavak hány szövegszót jelentenek. Az ötödik oszlop a szóalakokhoz viszonyított arányt, a hatodik pedig a teljes szövegszóhoz viszonyított arányt mutatja. Ha közelebbről megnézzük, tanulságos, hogy a 27 ezres szókészletből több mint 17 ezer szó, kb. 64%, csak egyszer fordul elő. Természetesen itt a toldalékos alakokat külön számítottuk.

A projektstatisztikában nem kapunk túl sok plusz információt, és ez is inkább technikai jellegű.

Analysis based on the whole vocabulary	Az elemzés a teljes szöszedetre épül
Total vocabulary = 27068 types	Teljes szókészlet = 27068 szóalak (típus)
Project wordcount = 142757 tokens	Projekt szószáma = 142757 szó
Types/tokens = 0,18960892	Szóalak/összes szó = 0,18960892
Types/sqrt(tokens) = 71,64031082	Szóalak/
Yule's k = 154,35324502	Yule

49. ábra: A projektstatisztika adatai

Az utolsó rész betűgyakoriságot mutat. Ebből tudhatjuk meg, hogy az egyes betűk hányszor szerepelnek a teljes szövegben. Talán még egy fontos dolgot kell megemlítenünk. A 47. és 48. ábrán levő ikonsor alatt láthatunk egy KeyWords feliratú gombot. Ha erre kattintunk, a következő ablakot láthatjuk:



50. ábra: Kulcsszavak szerkesztése

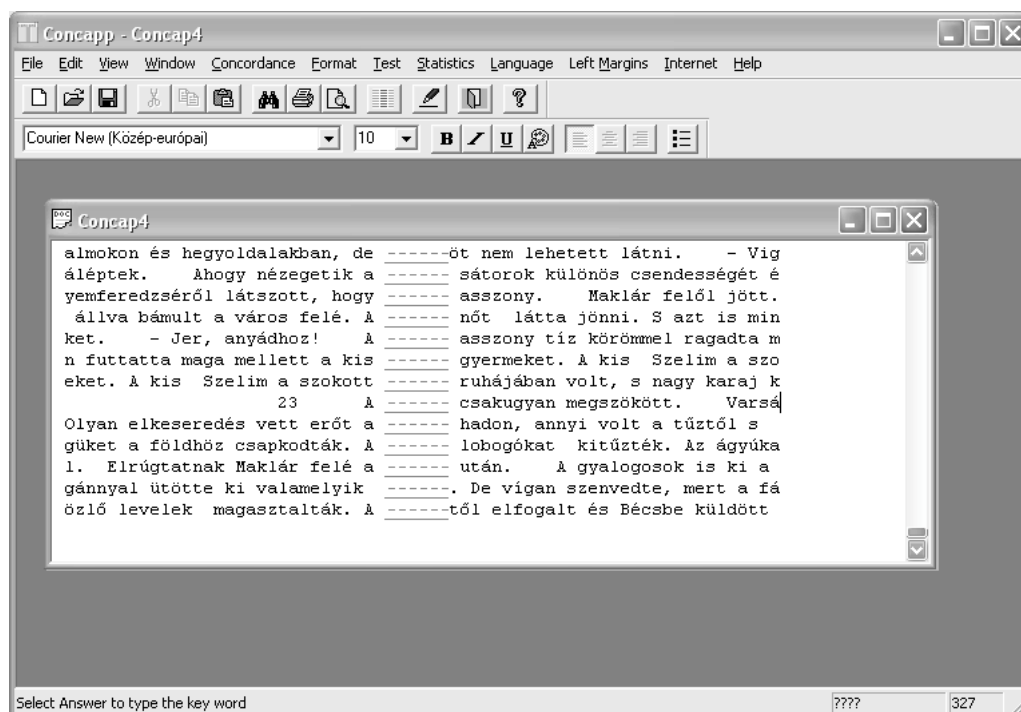
Az ablak két jól elkülöníthető, bal és jobb oldali listát tartalmaz. A bal oldalon látható az összes egység, amelyet a számítógép szóként értelmez. A szó meghatározása a számítógép számára az, amit két oldalról szóköz határol. Jól látható a bal oldali oszlopban, hogy a gondolatjelet és a számokat is szónak tekinti a program. Ha azonban a felhasználó ezeket nem tekinti annak, akkor nincs rájuk szükség a szótárban. Az ilyen felesleges „szavak” szűrésére szolgál ez az ablak. Hozzá lehet adni, vagy törölni lehet szavakat. Ebben az esetben látható, hogy a jobb oldali oszlopban már nem szerepelnek a számok, töröltük őket. Automatikusan is törölhetők elemek, például a minimális gyakoriság megadásával. Az ablak jobb oldalán látható, hogy a kulcsszavakat exportálni és importálni is lehet.

A program menüi meglehetősen egyszerűek. Talán a Tools menü Apply a Stop List almenüjéről érdemes még szólnunk. Mi is a *stop list*? Minden nyelvben vannak olyan szavak, amelyek nagyon gyakran előfordulnak, de igazából nem vagyunk rájuk kíváncsiak vizsgálatuk során, mert nem jelentéshordozók. Az angolban ilyen például a névelő, a prepozíciók stb. Annak érdekében, hogy ezek ne „szennyezzék” a listánkat, egyszerűen kikapcsolhatók, tehát nem jelennek meg a gyakoriság vizsgálatakor, ha ezt úgy kívánjuk.

4.4.3. ConcApp

E program indítása után arra vigyázzunk, hogy ne MS-DOS-os szövegfájlként, hanem DOS dokumentumként nyissuk ki a vizsgálni kívánt fájlt. Ez a program angol, kínai és japán szövegek vizsgálata céljából készült, így bizonyos funkciók nem működnek ideálisan a magyar nyelv esetében. Ezt a programot nem írjuk le olyan részletességgel, mint az előző kettőt, hanem olyan funkcióját említjük elsősorban, amely a többiben nem található meg.

A program indítása után megjelenik egy információs ablak, amely kattintásra eltűnik. A vizsgálandó fájl kiválasztása után a menüsorból válasszuk ki a Test menüpontot. A New választásával új ablak jelenik meg, ahol begépelhetjük a kívánt kifejezést vagy szót. Az OK gombra való kattintással a következő ablakot kapjuk:



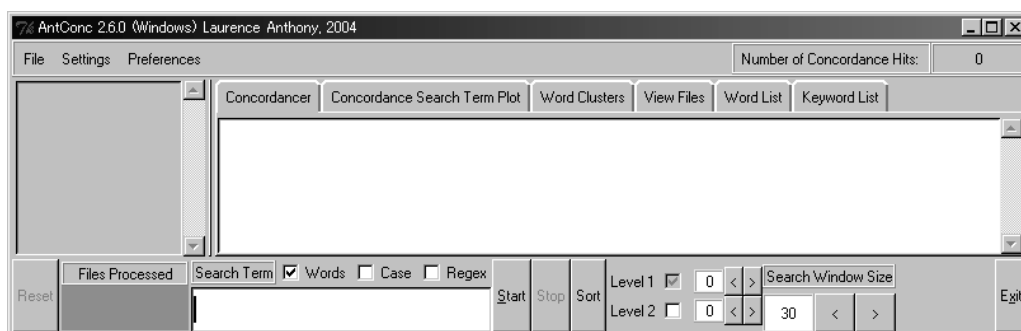
51. ábra: Teszt funkció a *ConcApp* programban

A keresett szót vonal helyettesíti, ami arra ad alkalmat, hogy a diák a keresett szót kitalálja. E funkció segítségével könnyen lehet játékos tesztek készíteni. Ezt a funkciót találtuk a leghasznosabbnak ez esetben.

4.4.4. AntConc

Ez a program sajnos nem képes a magyar ékezetes betűket értelmezni, ha szavakat keresünk. Így tehát szólistát nem tudunk ezzel készíteni. Ennek ellenére vannak olyan funkciók, amelyeket most is jól lehet használni. Mivel azonban a programok állandó fejlesztés alatt állnak, elképzelhető hogy mire e könyv az olvasó kezébe kerül, ez a probléma megoldódik, és így az itt leírtak nem fedik majd teljesen a valóságot. (Ez lenne a jobbik eset.)

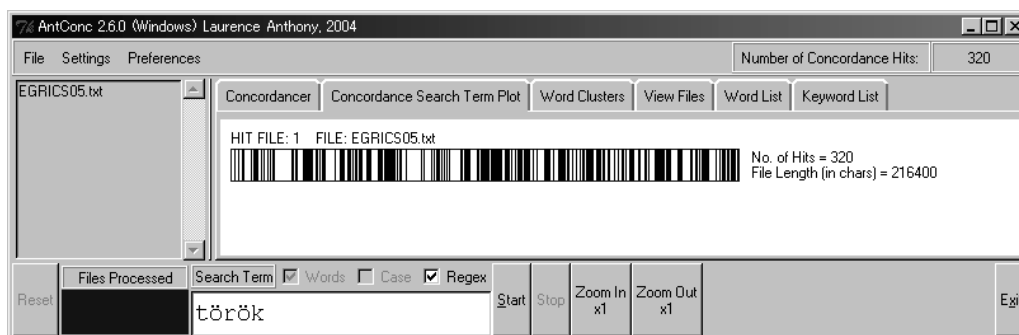
A program kezelőfelülete jól áttekinthető (52. ábra), a keresett szót vagy kifejezést az alsó szövegablakba kell beírni. A kurzor a program indításakor automatikusan ott villog, így nehéz eltéveszteni. Közvetlenül e fölött található a keresés módja, ahol a szóra való keresést vagy a regexet jelölhetnénk be, ha a program kezelni tudná az ékezetes betűket. Így csak a regex vezet eredményre. Természetesen a keresés megkezdéséhez meg kell nyitnunk egy fájlt, amit a szokásos módon a fájl menüből tehetünk meg.



52. ábra: Az *AntConc* kezelőfelülete

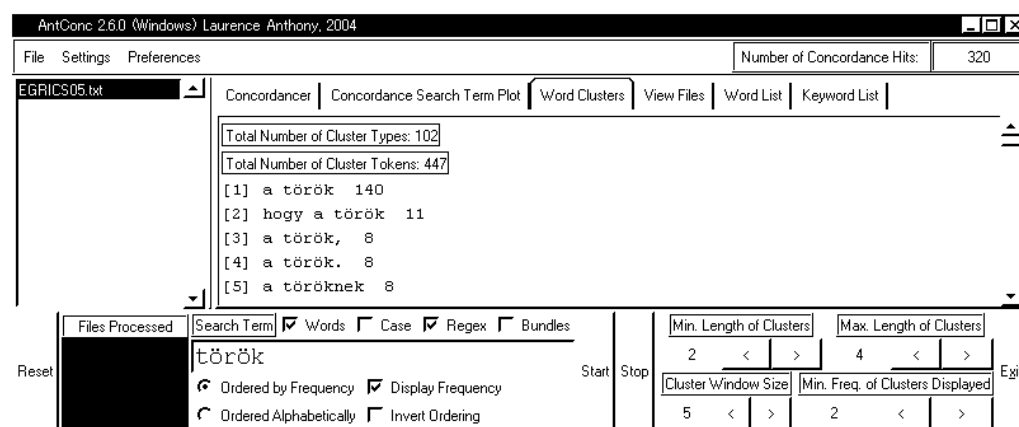
A keresett kifejezés konkordanciái a középső, nagy szövegablakban jelennek meg, a keresett kifejezés kék színnel van jelölve. A konkordanciákat a Level 1 és Level 2 gomb melletti szám, a képen 0, megváltoztatásának segítségével lehet úgy rendezni, hogy a különböző szerkezetek könnyen észrevehetőek legyenek. Ha például az 1R-t választjuk és a Sort gombra kattintunk, akkor a keresett szó melletti jobbra eső szavakat fogja ábécé szerinti sorba rendezni. Nincs megszabva, hogy hányas számra állíthatjuk ezt a funkciót, de két vagy háromnál többre nemigen érdemes állítani. A képen 30-ra állított szám változtatásával az ablakban egy sorban szereplő szövegmennyiséget lehet változtatni.

A keresés eredményét megtartva, a Concordance melletti fülre kattintva a vásárlásból jól ismert vonalkódra hasonlító képet láthatunk (53. ábra), mely azt szemlélteti, hogy a keresett kifejezés a szövegben hol helyezkedik el. Minden egyes előfordulást egy függőleges vonal jelez, így a sűrű vonalak azt mutatják, hogy ezek egymáshoz igen közel vannak.



53. ábra: A keresett szó elhelyezkedése a szövegben

A Word Clusters fülre való kattintással újabb formában vizsgálhatjuk a kívánt szót vagy kifejezést. Ne feledjük azonban, hogy most is csak regexként kereshetünk. Mire jó ez a funkció? Az MLCT program leírásakor említettük az n-gram keresési módját, és hogy ez meglehetősen igénybe veszi a számítógép kapacitását. Az AntConc e funkciója arra használható, hogy a keresett kifejezést az általunk megadott minimum és maximum szócsoportokban kigyűjtse előfordulásuk számával együtt.



54. ábra: Szócsoport-keresés

A keresett szó alatti részen lehet beállítani, hogy kívánjuk-e a gyakoriságot a képernyőn látni vagy sem, hogy milyen sorrendben, növekvő vagy csökkenő gyakoriság szerint, ábécé szerinti vagy fordított sorrendben akarjuk-e látni a szócsoportokat.

A View Files az eredeti szövegfájlt mutatja, a keresett szóra lehet a szövegben ugrani az előző (Previous Hit) és a következő (Next Hit) gombok segítségével. Itt meg szeretnénk jegyezni, hogy a konkordanciák nézetből is a szövegre lehet ugrani, ha a keresett szó fölé vitt egérmutató átváltozik kézzé, és ekkor a szóra kattintunk.

A Word List, azaz szólista funkció használhatatlan a magyar nyelv esetében, ha ékezetes betűk is szerepelnek a szövegben.

A Keyword List használatához egy referenciakorpuszra is szükség van. Mivel a program a magyar ékezetes szavakat nem értelmezte szavakként, ezt a funkciót ki sem próbáltuk.

4.5. Összefoglalás

Manapság nem az információhiány a probléma, hanem inkább az, hogy a rengeteg rendelkezésre álló, bennünket időnként elöntő áradatot nehéz befogadnunk. A számítógépek fejlődésével és az internet elterjedésével ez nem csak a magánéletre igaz. Az internet és az elektronikus könyvek korában a nyelvi és nyelvészeti vizsgálatokhoz rendelkezésre álló adathalmaz hatalmas. Minden országban igyekeznek minél nagyobb nemzeti korpuszokat létrehozni a korpusz alapú nyelvészeti leírások érdekében. Ebben a fejezetben először a korpuszkészítéskor használt programokat említettük meg, de nem írtuk le ezeket részletesen, hiszen elsősorban nem az annotált korpusz készítésére akarjuk biztatni az olvasót, hanem már „kész” korpuszok használatára, vagy egy saját szövegfájlokból álló korpusz használatára.

A korpuszok elemzésének egyik alapvető módja a konkordanciák elemzése. A konkordanciaprogramok általános működésének bemutatása után röviden áttekintettük a korai konkordanciaprogramokat, amelyek még ma is elérhetők az interneten.

Az internetes felületen futó konkordanciaprogramok a kimondottan magyar nyelvre készültek kivételével nem használhatók jól a magyar nyelv esetében az ékezetes betűk miatt. Így a fejezet nagy részét annak szenteltük, hogy négy, az internetről ingyenesen letölthető programot több-kevesebb részletességgel bemutassunk (MLCT, SCP, ConcApp és AntConc). Azért éreztük szükségét annak, hogy több programot is bemutassunk, mert egyikük sem a magyar nyelv speciális igényeinek figyelembevételével készült. Így a különböző programok különböző funkciói egymástól eltérő módon működnek, alkalmasabbak vagy kevésbé alkalmasak a magyar szövegek vizsgálatára.

Javasoltuk, hogy a programleírások olvasásával egy időben az olvasó is próbálja ki az adott program működését a saját számítógépén, saját szövegfájlja segítségével. A következő fejezetben utalunk a programok különböző funkcióira, de végrehajtásuk módját nem ismertetjük újból.