Alkalmazott Nyelvtudomány
VIII. évfolyam 1-2. szám 2008.

LESI ZOLTÁN

Pázmány Péter Katolikus Egyetem, Informatika Kar

zoli@nix.hu

Automatikus formai verselemzés

In this article we shortly present the field of computational poem analysis and our results. An automatic poem analyzer can ease the linguists' work, particularly when a large corpus is analyzed. This tool can help to prove statements in literature and linguistics. We analyzed sonnets and sonnet translations of the world-famous poet, Sándor Weöres. We consider his work as a part of the Hungarian and the international cultural heritage. Weöres' collected writings are digitally available, thanks to a provisional philological study. The exceptional researcher, Iván Fónagy's clear aspects served as a basis of our work. Phonetical description is essential for prosodic analysis. We studied in details the computational processing of metrics, alliterations, rhymes for which we applied learning algorithms. For the representation of our results we used the TEI P4 XML format. The analysis and the received statistical information can help recognize many hidden features of Weöres' sonnets.

Bevezetés

A szépirodalmi szövegeket többféle szempont szerint csoportosíthatjuk. Horváth Iván (1990) *A vers* című könyvében három megközelítést mutat be. Ha a transzcendens versolvasó szemüveget tesszük fel, a verset nem tudjuk másnak látni, mint költeménynek, ilyenkor a verselmélet egybeolvad a költészet elméletével. Tekinthetjük a verset nyelvi egyetemességnek: olyan megjelölt beszédmódnak, amely valamiképpen minden természetes nyelvben létezik. A harmadik módszer szerint a vers az, amit egy bizonyos irodalmi hagyomány részesei annak tartanak.

A számítógépes nyelvészetben a verselemzés új kutatási terület, hiszen magyar nyelvű szövegekre kidolgozott (vagy magyar szerző munkájaként ismert) *automatikus* verselemző programról nincs tudomásunk. A probléma megoldása talán azért is váratott magára, mert nem könnyű elhatárolni a megoldható és egyelőre megoldhatatlannak tűnő részeket.

Weöres Sándor és a korpusz

Weöres Sándor (1913-1989) költő, műfordító, író verseit elemeztük. Eleinte úgy tűnt, hogy Weöres lesz a második Nyugat-nemzedék költői reneszánszának rokokó feloldozója, de lírája hamarosan tágulni, mélyülni kezdett, „kozmikus értelemben” egyre vallásosabbá vált, telítődött tragikus érzületekkel. Bori Imre tanulmányában (Bori, 1984) írja, hogy jellemző Weöres költészetére egy fontos zenei mozzanat a kettősség: ha Weöres verseinek nagyobbik hányadát a zene

fogalmával helyettesíthetjük, úgy van egy verscsoportja, amely a „nem-zenét” jelenti. Ez a zenei elv összefogja Weöres költőiségének alapvető törekvését, amely a harmónia utáni vágyban s a költői megvalósulásában nyilatkozik meg.

Weöres megkísérelte azt, amit rajta kívül senki: a magyar költészetben addig még meg nem honosított, a görög-római verselésben is ritkaságszámba menő különös ritmusegységekkel kísérletezett, önálló versalkotó tényezővé téve a helyettesítő lábakat.

A szonett klasszikus formáját Petrarca alakította ki, a strófatagolása: két négysoros és két háromsoros strófa (rímképlete: abab abab cde cde). Ma már szonettnek nevezhetnek minden tizennégy soros verset, akár szabadverset is. Weöres szerint (Weöres, 1977) „A szonett első nyolc sora a nyolcoldalú kristály, az oktaéder: a végső hat sor az előbbiek ismétlése, más összeállításban, más összefüggésekkel.”

Nagy L. János és Alexin Zoltán 1999 és 2002 között létrehozták a virtuális kritikai kiadás 'editio princeps'-ét, hogy minél teljesebb Weöres-korpuszon dolgozhassanak (Nagy és Alexin, 2004). Az 1986-ban megjelent háromkötetes *Egybegyűjtött írások* című gyűjteményt vették alapul.

Kapcsolódó munkák

A *Corpus Poeticarum* munkálatai Léon Robel és Roman Jakobson kezdeményezésére indultak el 1973-ban. Ez a korpusz minden nemzet minden verselési rendszerének teljes és kimerítő leírását gyűjti. A kezdeményezéshez kapcsolódik Horváth Iván (Horváth, 1999) vezetésével a szegedi fejlesztésű számítógépes adatbázis: *Répertoire de la poésie hongroise ancienne*. Ezek a nagyszabású projektek emberi elemzést alkalmaztak és az eredményeket adatbázisokban tárolták, tehát nem tekinthetők automatikusnak.

A *The Metrometer* verselemző eszköz (Beaudouin & Yvon, 1996) francia szövegeket (Corneille és Racin drámákat) elemzett. A rendszer tartalmaz egy modult, amely megadja a fonetikus átírást, felhasználva a versek szintaktikus elemzését. A Metrometer 80 000 verssoron, ami körülbelül 70 000 szó, majdnem tökéletes metrikai eredményt adott. A Metrometer kizárólag a metrikára koncentrál, más verstani szempontokkal nem foglalkozik.

Az angol verseket elemző *Sound Patterns* eszköz 2004-ben készült (Love, 2004). A hangtani elemzés alapján kimutatja az angol szövegekben a végrímeket, alliterációkat, hangismétlődéseket, megszámolja a sorok szótagszámát, emellett foglalkozik az áthajlással, kérdő és felkiáltó mondatokkal is. Miután kötegelve beolvasta és elemezte a szövegeket, XML fájlban strukturálja az eredményeket. A *Sound Patterns* nyelvészeti és verstani alapjai bizonytalanok.

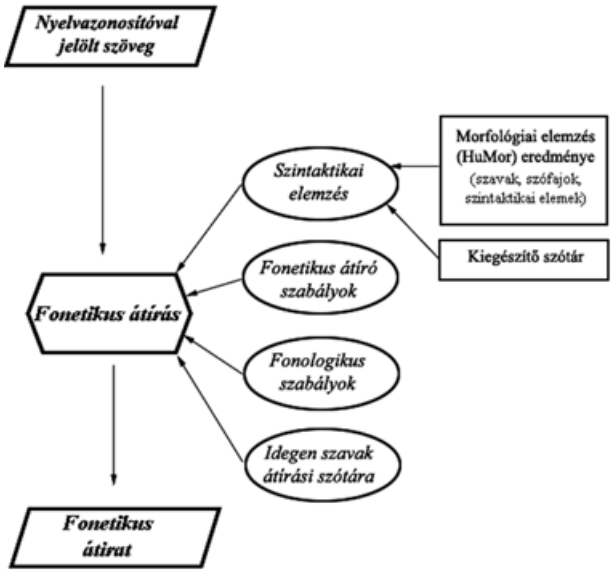
Fónagy Iván nemzetközi hírű nyelvész, pszichológus a hatvanas években megtervezetett egy programot, amely – formai megvilágításból – prózai és verses szövegekkel foglalkozik, majd kiegészítette két másik fejezettel: „Program köznyelvi szövegekre”, „Program költői szövegekre”. A köznyelvi szövegelem-

zést a következő részekre bontja: a fonémák, a fonemikus jegyek gyakorisága, a mássalhangzók egymásra hatása, a szótagszerkezet, a digrammák, a szavak, a mondatok hossza, a szófajok gyakorisága, a mondatrészek szófaja, a mondattípusok, a szórend, a felsorolás és a közbeékelt mondatok. A „Program költői szövegekre" kiegészítésben: áthajlás, egybecsengés (alliteráció, rím, asszonánc, alliteráló rím, kecskerím, mássalhangzós asszonánc), eufónia, metrum, ritmus, strófaszerkezet, rímképlet, címek témákat tárgyalja. A világos, statisztikai jellegű szempontokat tartalmazó tervek (Fónagy, 1997) adták a nyelvészeti és a verstani alapot programunkhoz.

Automatikus fonetikai elemzés

A legelterjedtebb és a legáltalánosabb érvényű fonetikus átírási rendszer az Association Phonétique Internationale (Nemzetközi Fonetikai Társaság) által 1889-ben elfogadott és APhi néven (ma inkább angol neve alapján IPA-ként) ismertté vált írásmód. Alapelvei pragmatikusak, az egyetemes használhatóság nem szempont, hanyagolja a fonetikai finomságokat, jellemző a nyomdatechnikai egyszerűség.

A fonetikus átíró modul működését az 1. ábra szemlélteti. A fonetikus átíró szabályokat Kassai Ilona (Kassai, 1998) táblázata alapján fogalmaztuk meg.



1. ábra

Az automatikus fonetikai elemző működése

A karakter környezetét megvizsgálva, a szintaktikai elemzés alapján döntjük el, hogy az adott helyen digramma (cs, dz, dzs, gy, ly, ny, sz, ty, zs) van-e.

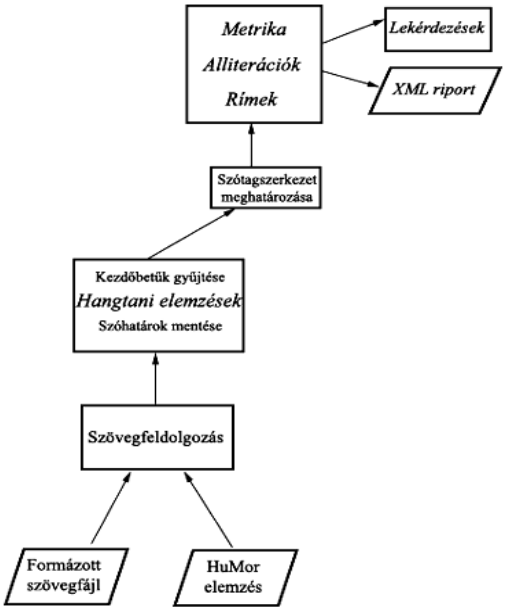
Amennyiben egy magyar szóról van szó, akkor a fonologikus szabályokat, idegen szó esetén az „idegen szavak átírási szótárát" vesszük figyelembe. Az idegen szavak felismerését egy nyelvazonosító rendszer (Kiss és Németh, 2006) alkalmazásával oldottuk meg.

Bizonyos hangoknak több ejtésváltozata van, így néhány újabb szabállyal kellett bővíteni a rendszert. A különböző típusú szóvégi *h*-k (pl. *méh, doh*) miatt a szóvégmutató szótárt (Papp, 1966) alkalmaztuk.

A hangok egymásra hatásakor megváltozik a fonetikus leírás, ilyenkor fonologikus szabályokat (pl. zöngésség szerinti hasonulást) alkalmazunk.

Az automatikus verstani elemző

Az automatikus verselemző működését a 2. ábra mutatja be. A heurisztikus elemző a már meglévő fonetikai elemzésen túl a fónagyi tervekre, valamint Szepes Erika és Szerdahelyi István (1981) *Verstan* című kötetére támaszkodik. Az időmértékes verselés legáltalánosabb jellemvonása az, hogy képleteiben a szótagok időértéke a legfontosabb elem. Az alliteráció a hangfestés egyik sajátos esete, amikor a szövegben egy-egy hang gyakorisága úgy nő meg, a szóban forgó hang a szavak elején állva ismétlődik. Alliterál a szó tágabb értelmében, az egy soron belül azonos mássalhangzóval vagy magánhangzóval kezdődő két szó. Szorosabb értelemben vett alliteráció esetében eltekintünk a névelőktől és kötőszavaktól. A félalliteráció elemzésében az alliteráló sorokban és azok környezetében megvizsgálandó, hogy nem alliterál-e hasonló hang; magánhangzóknál csak elsőfokú, mássalhangzónál maximum ötödfokú hasonlóság veendő figyelembe. A rím a szóvégi hangok egybecsengése a szövegben. További feltétel, hogy egymáshoz olyan közeli vagy olyan elhelyezésű szavakat kössön össze, amelyeknek egybecsengése világosan érzékelhető.



2. ábra

Az automatikus verstani elemző működése

A morfológiai elemzést (melyet a versek HuMor (Prószéky & Kis, 1999) elemzéséből nyerünk) és a hangtani vizsgálatokat felhasználva létrehozhatóak a számunka fontos alkalmazások a metrika, az alliterációk és a rímek. A megalapozott végeredményeket egy XML-fájlban összegezzük, és lekérdezéseket hajtunk végre.

A sorvégi egybecsengések illetve alliterációk vizsgálatához szükség van a hangok összehasonlítására. A fónagyi tervezet tartalmaz egy leírást a fonémák eltérési fokáról, mely egyszerűen algoritmizálható.

A metrika, az alliteráció és a sorvégi egybecsengések alkalmazásait heurisztikus és tanuló algoritmusokkal is meghatároztuk, majd a részeredményeket szavazással egyesítettük.

A metrika tanításához két jellemzőt kell megadni:

- a vizsgált szótag magánhangzóját,
- a magánhangzót követő mássalhangzók számát.

Az alliterációk, félalliterációk tanításához három jellemzőt adunk meg:

- a két kezdőbetű fonetikus eltérésének fokát,
- a két szó közül valamelyik kötőszó vagy névelő vagy sem,
- a szövegkörnyezet tartalmaz valamelyik kezdőbetűnek megfelelő tiszta alliterációt.

A sorvégi egybecsengés tanításához öt jellemzőt kell megadni:

- a két szótag magánhangzóinak eltérési fokát,
- a mássalhangzók minimális eltérési fokának összegét,

- a két szótag szerkezetének eltérési fokát,
- a következő szótagpáros egybecseng, illetve a vizsgált szótag sorvég vagy egyik feltétel sem teljesül,
- a tanulóhalmazban az olyan esetek számát, amikor a vizsgált két sor rímel.

Mindhárom alkalmazáshoz független tanuló- és teszthalmazra volt szükség, ezért a 192 versből álló korpuszt két részre bontottuk. A tanulóhalmazba a fent részletezett tulajdonságok alapján 41 vers elemzése, a teszthalmazba 151 vers elemzése került, ezen belül külön megvizsgáltuk az *Átváltozások* ciklust is. Talán kicsinek tűnik a korpusz mérete, de egy vers – mint az 1. táblázat mutatja – több száz példát ad mindhárom alkalmazás esetében, ez magyarázza az eredményeket is.

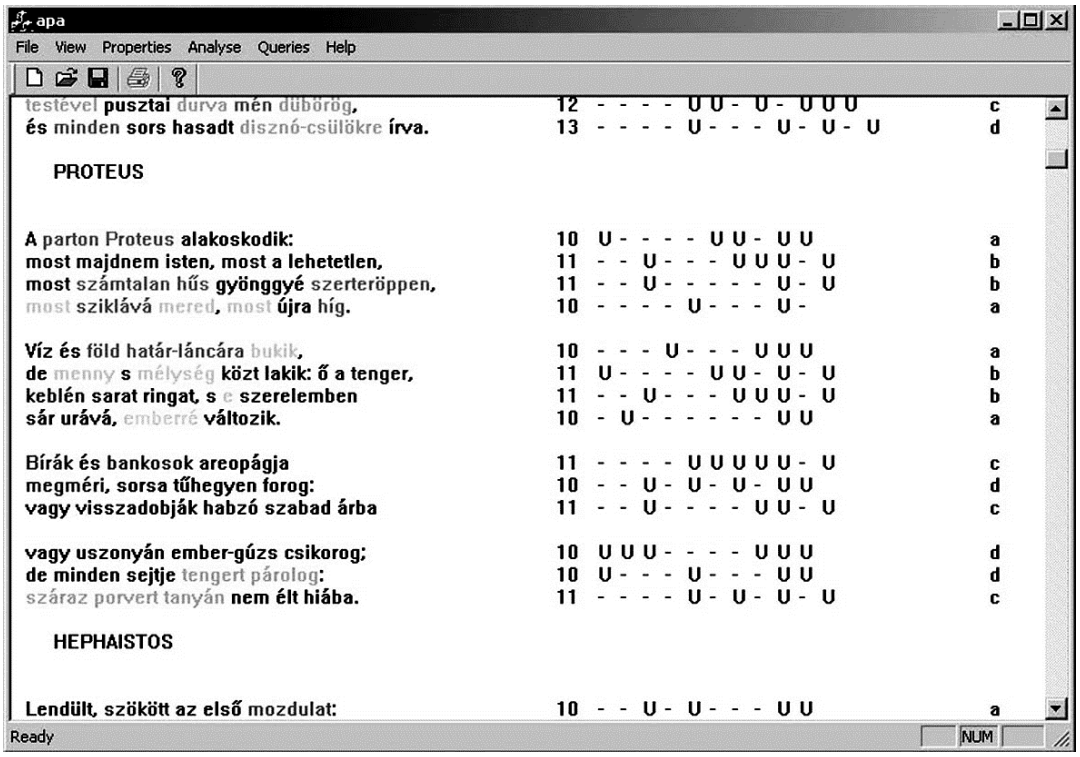
1. táblázat

A korpuszokban szereplő példák száma

Alkalmazás / Példák száma	**tanulóhalmaz**	**teszthalmaz**
metrika	5790	22380
alliteráció	19808	75530
sorvégi egybecsengés	7086	26916

Az elkészült formai verselemző-rendszer eredményei

A 2004 októberében elkezdett tudományos munka legfőbb eredménye, hogy a Fónagy-tervezet fejezetei alapján megterveztük és implementáltuk (C++ programozási nyelven) az első magyar automatikus verselemzőt (lásd 3. ábra). Az első időszak feladatai: a Fónagy-tervezet értelmezése, megvalósíthatósági tanulmány, DTD-tervezet készítése, valamint a verstani, hangtani ismeretanyag összegyűjtése voltak. A fónagyi tervezetben a statisztikai szempontokat visszavezettük alapadatokra. Nagy L. János kurzusán teszteltük a szempontrendszert, amely később a program alapjává vált. A program ellenpontjaként a diákok elemezték a verseket.



3. ábra

Egy elemzett szonett

Összegyűjtöttük és annotáltuk a virtuális kritikai kiadásból Weöres Sándor összes (101) szonettjét, valamint 91 szonett fordítását. Az annotáció csak a sort címet, alcímet, ajánlást, verssort jelöli. Ragaszkodtunk a szonettformához, bár korpuszunk így aránylag kis méretű lett, de mindenképpen homogén versformájú anyagot akartunk, mert így az egyes versekre adott eredmény könnyebben összevethető a korpuszéval, és a szonettformát is jellemzi.

A további vizsgálataink miatt pontos fonetikai eredmények szükségeltettek. A digrammák szétvágásához (pl. százszor) bevezettünk egy új módszert: az ilyen esetek az összetett szavak szóhatárainál jönnek létre, ezért a morfológiai elemzés alapján elvégezhető a vágás. Tesztjeink igazolták a módszer helyességét.

A magánhangzók eltérési fokának meghatározása hiányzott a Fónagy-tervezetből, ezt „A mássalhangzók eltérési foka” című fejezet alapján pótoltuk. A metrika, az alliterációk és a rímek heurisztikus meghatározása a Fónagy-tervezet alapján történt.

2. táblázat

A teljes korpuszra és az Átváltozások ciklusra vonatkozó eredmények

	Heurisztikák	**C4.5**	**SVM**	**Döntés után**
Metrika 151 szonettre	100.00%	100.00%	99.97%	100.00%
Alliteráció 151 szonettre	98.69%	98.66%	95.69%	99.80%
Alliteráció az *Átváltozások* ciklusra	98.53%	98.53%	95.95%	99.60%
Sorvégi egybecsengések 151 szonettre	79.14%	90.81%	89.19%	90.87%
Sorvégi egybecsengések az *Átváltozások* ciklusra	81.72%	94.80%	93.25%	94.38%

Az eredmények pontosításához a C4.5 és az SVM tanuló algoritmusokat használtuk.

A C4.5 (Quinlan, 1993) a döntési fák módszerén alapul, akkor alkalmazzuk, ha (attribútum-érték) párokkal ábrázolható példáink vannak. A példa egy adott attribútum-halmaz elemeiből, és az ezekhez tartozó értékekből áll. Előnyös, ha ezek az értékek egy kis elemszámú halmazból vesznek fel értéket; a célfüggvénynek diszkrét, lehetőleg bináris kimenete van (pozitív, negatív), de az algoritmus könnyen átalakítható több mint kétértékű kimenetre; a példák hibákat és hiányzó attribútumokat is tartalmazhatnak.

Az SVM (Support Vector Machine) egy eszköz adatok osztályozására (Chih-Chung Chang, Chih-Jen Lin, 2001). Az SVM célja, hogy létrehozzon egy modellt, amely megjósolja a teszt halmaz célértékeit. A tanulóvektort magasabb dimenziójú térben ábrázoljuk a ϕ függvénnyel. Az SVM keresi a lineáris hipersíkot, amely a maximális eltéréssel választja el a pontokat a magasabb dimenziójú térben.

Felhasználva a két tanuló algoritmust, létrehoztuk a három alkalmazás (metrika, alliteráció és sorvégi egybecsengések) modelljét, és dekomponáltuk a feladatokat. A rímeket szétbontottuk egy szótagos egybecsengésekre, ahol elsősorban a szótag tulajdonságait, valamint a tanulóhalmazban az egymással rímelő sorpárok gyakoriságát vizsgáltuk.

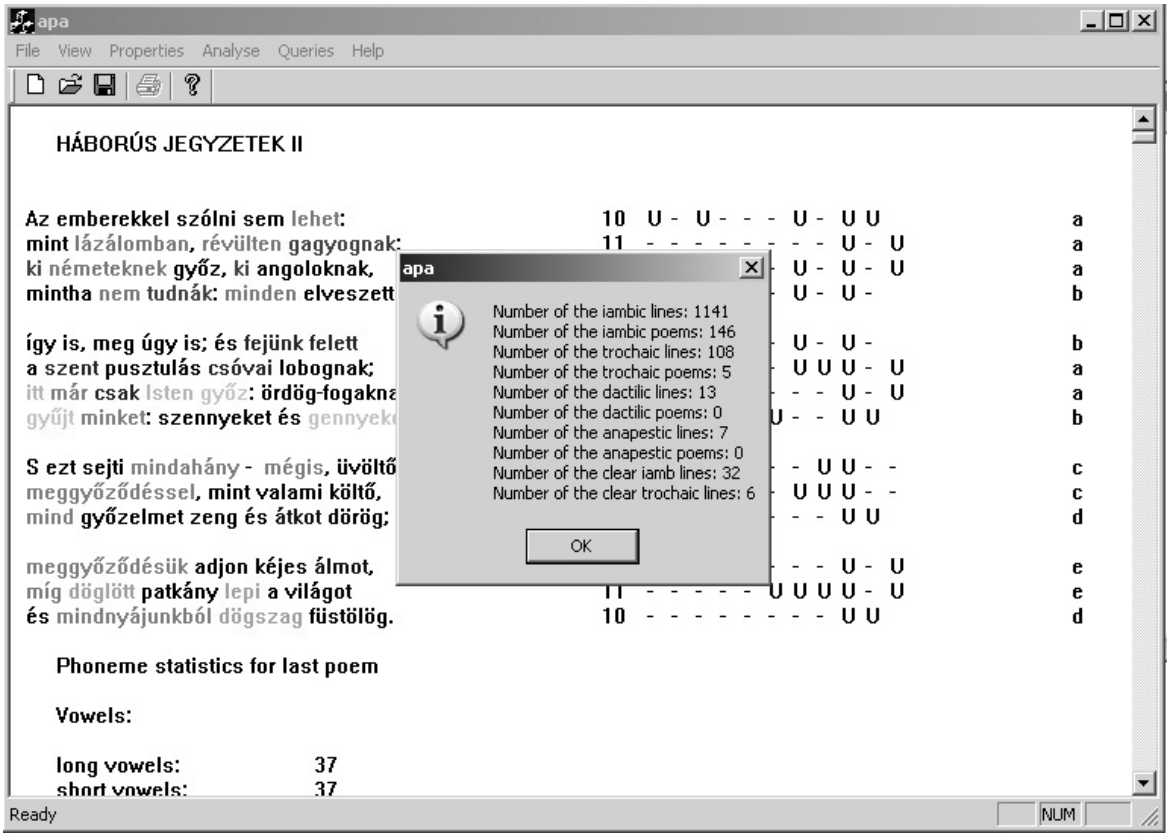
Az eredmények meghatározásához használt metrika az F-érték volt, alapja a releváns és a kitermelt verstani jellemzők (egyetlen szótag hosszúsága, két szókezdő hang egybecsengése és két sorvégi szótag egybecsengése) által meghatározott arányok.

A két tanuló algoritmus és a heurisztika eredményeit szavazással egyesítettük, a végeredmény (lásd 2. táblázat) az alliterációk vizsgálatakor pontosabb, a sorvégi egybecsengések esetén nem változott szignifikánsan, a metrikai alkalma-

zásban pedig mind a tanuló modellek, mind a heurisztikus algoritmusok kiváló eredményt adtak. Az eredményeink bizonyítják, hogy érdemes az alliterációk és rímek alkalmazásokra tanuló algoritmusokat használni. Meglepő eredmény, hogy a C4.5 pontosabban osztályozott, mint az SVM, ezt a kétszintű osztályozás és a diszkrét értékek okozzák.

A 2. táblázat a 152 szonettből álló tesztkorpuszra és az *Átváltozások* ciklusra lefuttatott sorvégi egybecsengéseket, alliterációkat vizsgáló alkalmazások eredményeit mutatja be. Az *Átváltozások* ciklusban lényegesen kevesebb egy szótagos asszonáncot találtunk, ennek köszönhető, hogy a kisebb korpuszban a rímek alkalmazás szignifikánsan jobban szerepelt.

Az elemzés elvégzése után – a Fónagy-tervezet alapján – fonetikát, metrikát, alliterációkat és sorvégi rímeket vizsgáló lekérdezéseket hajtottunk végre (lásd 4. ábra).



4. ábra

A metrikai elemzés eredménye 101 szonettre

A végeredményeket egy nemzetközileg elismert formátumban, egy TEI kompatibilis XML-fájlban (lásd 5. ábra) összegeztük (Sperberg-McQueen & Burnard, 2004), ehhez felhasználtuk a már kész DTD-terveket. A TEI ajánlást betartva, a minimális, szükséges módosítással oldottuk meg a konverziót.

```
- <w phon="kɔntsɔtɛj" syll="CVCCVCVC">
    <seg id="ALLT10404"> kancatej</seg>
    <ana>FN+FN</ana>
  </w>
```

5. ábra

Részlet az elkészült XML-ből

Hibaelemzés és a számítógépes verselemzés távlatai

A fonetikai rendszer legnagyobb pontatlansága az idegen szavak átírásából ered. Hiba már a nyelvazonosításnál felléphet. Hibás fonetikus átírás esetén, a fonemikus jegyekre, szótag-szerkezetekre, digrammákra, szóhosszúságokra, rímekre, alliterációkra és metrikára pontatlan eredményeket kapunk.

A szintaktikai elemzés rohamosan fejlődik, de pontatlansága problémát okoz a szófajok, alliterációk vizsgálatakor. A Fónagy-tervezetben szereplő szempontok automatizálhatóak, azonban többértelműségek miatt pontatlanságra számíthatunk. Például: a központozás hiánya bizonytalanná teheti a tagmondatokra bontást, a felsorolások, közbeékelt mondatok felismerését, a mondatok hosszát és a modalitását.

A versekben nem jelölt metrikai kétértelműségek elemzése a gép számára megoldhatatlan. Pl. „A mint B” (W.S: *Önéletrajz*) az „A” hosszan ejtendő, de ezt nem jelöli semmi a gép számára. A versrendszer (időmértékes, ütemhangsúlyos vagy szimultán) felismerése is probléma lehet a gépi elemzés számára – mivel megfelel a szabályoknak – az *Átváltozások* ciklusban szereplő *A nyüzsgés* című szonettet jambikusnak vesszük, de valójában nem csak időmértékes.

A rímképletek meghatározásakor a legnagyobb pontatlanságot a következő eset okozza: a két sorvég ugyan egybecseng, de ez a képletben nem jelenik meg. A gyakran előforduló rímképletek felismerésével javíthatjuk a *gyenge* sorvégi egybecsengések szűrését.

Új irányokat, problémákat fedezhetünk fel a számítógépes verselemzés kutatási területén belül, emellett eredményeink pontossága lényegesen növekedne, ha bővítenénk a korpuszunkat. Más költők műveinek elemezésével – a különbségek alapján –, a költők által használt verstani jegyek meghatározása egyszerűvé válna. Érdekes lenne összehasonlítani a régi magyar vers repertóriuma

formai eredményeit az automatikus elemző eredményeivel. Egyelőre beláthatlannak tűnik a rendszer más nyelvekre való alkalmazása, mert a fonetikai rendszer mellett a verstani alapok (magyar nyelv esetén lásd Fónagy 1997) is teljesen különböznek.

A formai jegyek szerepét, illetve a rímek és az alliterációk jelentőségét – a mű értelmezésének tükrében –, nem tudjuk számítógéppel meghatározni.

***

Köszönetnyilvánítás

Nagy elismeréssel tartozunk Fónagy Ivánnak, akinek a gondolatai meghatározóak voltak munkánk során. Szeretnénk köszönetet mondani a Szegedi Tudományegyetem tanárainak, dr. Alexin Zoltánnak és dr. Nagy Jánosnak a kutatás folyamán nyújtott értékes segítségért és a Budapesti Műszaki Egyetem kutatójának, Kiss Gézának a nyelvazonosítóért.

Irodalom

Bori I. (1984) A szintézisteremtő. *Bori Imre huszonöt tanulmánya a XX. Századi magyar irodalomról*. Újvidék: Forum Kiadó.

Beaudouin, V. and Yvon, F. (1996) The Metrometer: a Tool for Analysing French Verse. *Literary and Linguistic Computing*, 11/1.

Chih-Chung Chang, Chih-Jen Lin. (2001) *LIBSVM: a library for support vector machines*, (http://www.csie.ntu.edu.tw/~cjlin/libsvm)

Fónagy I. (1997) *Program prózai és verses szövegek elemzéséhez.* Antony. Kézirat.

Horváth I. (1990) *A Vers*. Budapest: Gondolat Kiadó.

Horváth I. (1979-1999) *A régi magyar vers repertóriuma* (Repertoire de la poésie hongroise ancienne). (http://magyar-irodalom.elte.hu/cgi-bin/repertorium/kezdolap?lan=eng)

Kassai I. (1998) *Fonetika*. Budapest: Nemzeti Tankönyvkiadó.

Kiss G. és Németh G. (2006) Machine-learning Algorithm for Automatic Labelling and its Application in Text-To-Speech Conversion. *Híradástechnika*, 2006/3. 51-58.

Love, T. (2004) Analysing Sound Patterns [draft] (http://www2.eng.cam.ac.uk/~tpl/asp/)

Nagy L. J. és Alexin Z. (2004) Weöres költői nyelvének számítógépes feldolgozása. In: Alexin Z. és Csendes D. (szerk.) *II. Magyar Számítógépes Nyelvészeti Konferencia*. SZTE TTK Informatikai Tanszékcsoport, Szeged.

Papp F. (1966) *Szóvégmutató szótár*. Budapest: Akadémiai Kiadó.

Prószéky, G and Kis, B. (1999) A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. College Park, Maryland, USA: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. pp. 261-268.

Quinlan, J. R. (1993) *Programs for Machine Learning*. San Meteo, CA: Morgan Kaufmann.

Sperberg-McQueen, C. M. and Burnard, L. (2004) TEI P4 Guidelines for Electronic Text Encoding and Interchange (http://www.tei-c.org)

Szepes E. és Szerdahelyi I. (1981) *Verstan*. Budapest: Gondolat Kiadó.

Weöres S. (1977) Oktaéder-kristály. *Új Írás*, 1977/4.