# Panaszlevelek szerkezetének gépi felismerése

Bártházi Eszter[1,2⋆], Héder Mihály[3,4]

[1] MTA SZTAKI Géppel Támogatott Megértés Kutatócsoport, barthazi@sztaki.hu
[2] SZTE BTK Nyelvtudományi Doktori Iskola, Elméleti Nyelvészet Program
[3] MTA SZTAKI Internet Technológiák és Alkalmazások Központ, mihaly.heder@sztaki.hu
[4] Budapesti Műszaki és Gazdaságtudományi Egyetem Filozófia és Tudománytörténet Tanszék

## 1. Bevezetés

Kutatásunk célja egy olyan rendszer fejlesztése, amely egy adott intézmény ügyfélszolgálatára beérkezett panaszlevelek megbízható tartalmi kivonatolására képes, majd ennek továbbfejlesztéseképpen dialógus formájában képes segítségére lenni a panaszos ügyfeleknek. Természetesen nem célunk az ügyfél–ügyintéző kapcsolatot pusztán ember–gép kommunikációra korlátozni, ennek megvalósíthatósága amúgy is kétséges, a cél az, hogy az ügyintézők dolgát segítendő a gép olyan mértékű segítséget nyújtson az ügyfélnek, amelyben még biztonsággal kompetensnek tekinthető.

A hivatalos levelezésnek megvan a maga formátuma, vannak elvárások a szerkezetére, tartalmára, valamint a szókincsére vonatkozóan. A korpuszt alkotó, az Igazságügyi és Rendészeti Minisztérium ügyfélszolgálatára érkezett panaszlevelek[5] azonban olyan szabadon alkotott dokumentumok, amelyek írásakor a levélírók az esetek nagy részében nem követték a megalkotásukra vonatkozó szokásos „előírásokat". Ennek megfelelően a levélírók leveleikben nem kizárólag a megoldásra váró problémájukra szorítkoznak, a leveleket hétköznapi nyelvhasználat, a szakkifejezések rendszertelen és pontatlan használata, valamint zavaros megfogalmazás jellemzi. A zavaros megfogalmazás, valamint a többletinformációk a gép számára inkonzisztens információt jelentenek, amely jelentősen megnehezíti, ha nem ellehetetleníti a levelek megfelelő tartalmi kivonatolását. A humán nyelvhasználó számára ez nem jelent problémát, hiszen korábbi tapasztalataira hagyatkozva, valamint a diskurzuskontextus feldolgozásával képes megfelelő kontextust építeni az információk értelmezéséhez. Azt szeretnénk, ha a gép is képes lenne rendszerezni ezeket az inkonzisztens információkat azáltal, hogy kontextust, és egyúttal egyfajta interpretációt tudna rendelni hozzájuk. Ennek érdekében a panaszleveleket kisebb részekre, úgynevezett szerkezeti egységekre osztottuk, úgymint Bemutatkozás, Probléma, Lezárás stb. A szerkezeti egységek

⋆ Ezúton szeretném kifejezni hálás köszönetemet Németh T. Enikőnek és Vámos Tibornak a tanulmányhoz fűzött értékes megjegyzéseikért és támogatásukért.
[5] A korpuszért külön köszönet illeti dr. Vörös Editet, az Igazságügyi és Rendészeti Minisztérium Társadalmi Kapcsolatok Osztályának vezetőjét.

a bennük előforduló nyelvi információk kontextusaként szolgálnak, a tartalmi kivonatolás pontosságát pedig azáltal növelik, hogy jelölik, hogy milyen típusú információt hol érdemes keresni a levelekben, például a levélírót azonosító adatokat a Bemutatkozásban stb. Jelen tanulmányban azt szeretnénk bemutatni, hogy a gép számára milyen felszíni jellemzők állnak rendelkezésre, amelyek alapján a szerkezeti egységeket felismerheti.

A tanulmánnyal ugyanakkor arra is szeretnénk felhívni a figyelmet, hogy az emberek által szabadon alkotott, szerkezeti megkötésektől mentes dokumentumok kivonatolása esetén a szózsákmodell puszta alkalmazása feltehetőleg nem mindig vezet megfelelő eredményre.

A tanulmány felépítése a következő. A Bevezetést követően a 2. pontban bemutatjuk a szerkezeti egységeket, ezt követően pedig a 3. pontban azokat a jellemzőket, amelyeket a szerkezeti egységek azonosítása során figyelembe vettünk, továbbá az azonosítás sikerességének mértékéről számolunk be, valamint arról, hogy milyen elképzelésünk van még az eredmények további javítására, a hatékonyság növelésére. A 4. pontban összefoglaljuk a tanulmány eredményeit, végül a tanulmányt a Hivatkozások listája zárja.

## 2. A szerkezeti egységek

A levelek szerkesztésére vonatkozóan megfigyelhető, hogy az esetek nagy részében bár a levélíró feltehetőleg a legjobb tudása szerint fogalmazza meg panaszát, gyakran keveredik a hétköznapi és az egyes szakterületekre jellemző nyelvhasználat. A szakterminusok használata gyakran pontatlan, jelentése van, hogy eltér az adott szakterületen belül használatos jelentéstől. A levelek megfogalmazása zavaros, rendezetlen, és a levelek nagy hányadban tartalmaznak olyan információkat, amelyeknek a további érdemi ügyintézésben nincs szerepük. A levelek szerkezete sem egységes, sok esetben nem felel meg a hivatalos levelekkel szemben támasztott általános szerkezeti elvárásoknak. Ennek feltehetően a levélírók iskolázatlansága és alacsony társadalmi státusza az oka, bár erre vonatkozóan nem állnak adatok a rendelkezésre, mivel a korpuszt a kutatócsoport anonimizálva kapta.

A problémát az 1. ábra szemlélteti, amely egy panaszlevelet reprezentál. A panaszlevélben megfigyelhető, hogy a levél írója azt kéri, hogy értesítsék a Franciaországban élő nagybátyját megromlott egészségi állapotáról, ugyanakkor a levélben számos olyan téma is felmerül, amely a további intézkedések szempontjából irreleváns, mint ebben az esetben az arra vonatkozó információ, hogy az édesapja hogy bánt vele gyermekkorában, valamint, hogy az egyik fia a levélírás idején épp börtönbüntetését tölti.

Annak ellenére, hogy a levelek szerkezete nem egységes, mégis megfigyelhetőek olyan szerkezeti részek, amelyek a korpuszban következetesen visszatérnek, és amelyekből az egyes levelek felépülnek. A korpusz tanulmányozása alapján az alábbi tizenegy szerkezeti egységet szükséges megkülönböztetni:

1. **Megszólítás**: a levélíró azt fejezi ki valamilyen módon, hogy kinek szánja levelét, azaz kitől vár megoldást a problémájára, pl. Tisztelt [személynév/ti-

s 29. levél .
s Tisztelt Minisztérium A nevem Person édesanyám neve Name Apám Name volt .
s Az édesanyám itt halt meg Magyarországon 28 éves volt .
s Sajnos én genetikailag örököltem egy súlyos betegséget erős fájdalmaim vannak izom ízületi gyulladás mivel gyermekkorom óta depresszióban szenvedek ez súlyosbítja a betegséget .
s Az immunrendszerem elhagyott 2 éve klimaxolok ez felgyorsította a folyamatot .
s 67%-os rokkant vagyok egy középiskolás fiammal élek önkormányzati bérlakásban Gyulán .
s A másik fiam Name Gyulai börtönben van ez is nagyon elszomorítja a lelkemet .
s Nem sajnáltatni szeretném magamat csak az lenne a kérésem , hogy szóljanak a Francia Nagykövetségen ha tudnak segítsenek nekem .
s A bácsikámat Name értesítsék , hogy beteg vagyok .
s Én mostoha gyermeke vagyok ennek az országnak nem én akartam ide jönni apán hozott ide minket és nem engedte , hogy édesanyámmal visszamenjünk .
s És édesanyámnak emiatt megszakadott a szíve .
s A férjem 2000-ben rákban meghalt .
s Én mindig becsületesen éltem és sokat dolgoztam Angol Női szabó szakmámban és még betanított lakatosként is dolgoztam a Gyulai Mezőgépnél 10 évig .
s Kérem önöket legyenek szívesek megmondani nekik és sajnos nem beszélem a Francia nyelvet apám nem engedte a rossz gyermekkorom miatt nem is lett volna rá módom .
s Apám vadállat módra nevelt bennünket .
s Nekem az volt a legnagyobb bűnöm , édesanyámra nagyon hasonlítok .

1. ábra. Egy panaszlevél, amely az IM ügyfélszolgálatára érkezett

tulus/szervezet/stb.] (a szögletes zárójelben lévő kategóriák absztrakt címkéket jelölnek, amelyek a levelekben természetesen a konkrét beszédhelyzetnek megfelelő nyelvi elemekkel vannak kitöltve);

2. **Bemutatkozás**: a levélíró (ideális esetben) megadja mindazokat az adatokat, amelyek a kizárólagos azonosításához szükségesek, pl. Alulírott [személynév]. . . ;
3. **Cél**: a levélíró még a panasza ismertetése előtt röviden, néhány szóban, vagy egy mondatban kifejezi, hogy milyen területen vár segítséget, pl. Tárgy: nyugdíjügy;
4. **Probléma**: a levélíró azt a problémáját részletezi, amelynek kapcsán megoldást vár az IM részéről;
5. **Javaslat**: a Probléma alternatívája, erre abban az esetben találunk példát, ha a levélíró nem egy adott probléma megoldását várja a minisztériumtól, hanem ő maga tesz egy javaslatot valamivel kapcsolatban;

6. **Elismerés**: a levélíró elismerését fejezi ki a levél címzettjének eddigi tevékenységével, eredményeivel kapcsolatban, pl. Engedje meg, hogy gratuláljak...;
7. **Egyéb körülmények**: a levélíró a problémájához szorosan nem, vagy egyáltalán nem kapcsolódó egyéb problémáját, életkörülményeit, egészségügyi állapotát stb. ecseteli;
8. **Elvárás**: a levélíró azt fogalmazza meg, hogy milyen viselkedést, intézkedést vár el az ügyintéző részéről, pl. Kérem, hogy a fentiek alapján...;
9. **Köszönet**: a levélíró megköszöni az eddigi intézkedést, türelmet, illetve előre is megköszöni a további intézkedéseket, pl. Előre is köszönöm, hogy válaszlevelével megtisztel;
10. **Lezárás**: a levélíró egy adott formulával befejezi a levelét, pl. Minden jót!
11. **Csatolmányra hivatkozás**: a levélíró a levél egy mellékletére hivatkozik, pl. Mellékelten megküldöm kérelmemet.

A nyelvészeti pragmatikai, valamint a számítógépes pragmatikai kutatásokban egyre inkább az az uralkodó nézet, hogy a nyelvi és a kontextuális információk együttes figyelembevétele szükséges nem csak a megnyilatkozás-, de a szójelentés megalkotásában is (l. [1,2,3,4,5,6]). A panaszlevelek esetében ilyen kontextuális információ, hogy panaszlevélről van szó, amellyel egy magyar állampolgár az Igazságügyi Minisztériumhoz fordult, de a kellően pontos információkinyeréshez a fent említett, a panaszlevelekre jellemző tulajdonságok miatt ez még nem elegendő. A szerkezeti egységek kisebb, úgymond „minikontextusát" adják a bennük előforduló nyelvi kifejezéseknek.

A kontextus korábbi, statikus értelmezésével szemben annak dinamikus jellegét a pragmatikában először [1] fogalmazta meg. Eszerint a kontextust nem vehetjük előre adottnak, azt az értelmezés során kell felépíteni. A kontextus megnyilatkozásról megnyilatkozásra változik, valamint az egyes kifejezések is egymás kontextusaként szolgálnak az értelmezés során. Vannak továbbá olyan nyelvi mutatók, amelyek segítik a befogadót a megfelelő kontextus felépítésében, és ezáltal a megfelelő interpretáció megalkotásában. A szövegkörnyezet is segíti a kontextus felépítését, hiszen bizonyos kifejezések (együttes) jelenléte leszűkítheti az adott megnyilatkozás értelmezési lehetőségeit. Feltételezésünk szerint a szerkezeti egységek, azaz a „minikontextusok" felismerését ennek megfelelően bizonyos nyelvi kifejezések, azok nyelvtani tulajdonságai, valamint a szerkezetre vonatkozó heurisztikák segíthetik (l. 3. pont).

A szerkezeti egységek tematikus kontextusként szolgálnak a bennük előforduló nyelvi információ értelmezéséhez, azaz arra vonatkozóan szolgáltatnak háttértudást, hogy az adott nyelvi információ az adott kontextusban kire/mire vonatkozik, kiről/miről szól ([7]: 481). A szerkezeti egységek megmutatják, hogy a különböző, a levélíró panaszára, valamint elvárásaira vonatkozó lényegi információk a levél mely részében találhatók, egyszersmind lehetővé teszik a nem lényegi részek elhagyását.

A fent felsorolt szerkezeti egységek természetesen nem minden levélben fordulnak elő teljes repertoárjukban, és a sorrendjük is igen nagy változatosságot mutat az egyes levelekben, ugyanakkor a korpuszban újra vissza-visszatérnek.

A következő pontban a szerkezeti egységek azonosítására vonatkozó, általunk alkalmazott módszereket, valamint ezek eredményességét ismertetjük.

## 3. A szerkezeti egységek azonosítása: eredmények

A korpusz létrehozása a következőképpen történt: A teljes korpusz 888 panaszlevelet tartalmaz, amelynek 20%-án végeztünk kézi annotálást. Az így annotált 198 levélben jelöltük a szerkezeti egységeket, és absztrakt címkével láttuk el a következő entitásokat: SZERVEZET, TITULUS, SZEMÉLY, JOGHIVATKOZÁS és BETEGSÉG, valamint a helységneveket a rendszer automatikusan felcímkézte. Az annotálással összesen 1384 szerkezeti egységet kaptunk. Minden annotált szerkezeti egységet egy rövid dokumentumként kezelve tehát lett 1384 dokumentumunk, amelyek a szerkezeti egységeknek megfelelő tizenegy kategóriába lettek besorolva. A teljes korpuszt annotáltuk a Szegedi Tudományegyetem által fejlesztett magyarlánc [8] szoftverrel, amely a lemmákat és a POS-tageket adta meg. Egy további előfeldolgozási lépésként kiszűrtük a háromnál kevesebbszer előforduló szavakat és a szokásos stopszavakat is, majd az előfordulási értékeket tf-idf módszer szerint normáltuk. Végeredményként 2750 lemmát kaptunk.

A korpusznak többféle változatát készítettük el. Az első változat kizárólag lemmákat tartalmaz, a második szintén lemmákat tartalmaz, azzal a kiegészítéssel, hogy az igék lemmái el vannak látva igeidő (jelen vs. múlt), valamint igemód (kijelentő vs. felszólító) címkékkel. A harmadik változat szintúgy lemmákat tartalmaz, kivéve, hogy azok helyett a kifejezések helyett, amelyeket az annotálás során absztrakt címkével láttunk el, azok absztrakt címkéje szerepel. A következő kategorizáló megoldásokat alkalmaztuk:

- Döntési fák: itt a J48 és a REPTree algoritmusok szerint C4.5 [9] típusú döntési fákat hoztunk létre.
- Naive Bayes [10] kategorizáló
- SMO [11] kategorizáló

A tanításhoz általában véletlenszerűen kiválasztottuk az adatok 2/3-át, majd a maradék 1/3-dal teszteltünk, néhol azonban tízfordulós keresztvalidációt is végeztünk. A tanításokat a weka [12] keretrendszerrel végeztük.

### 3.1. Lexikai jellemzők

A kiindulási kísérletünk a korpusz tanítása volt lemmák szerint, minden egyéb információ hozzáadása nélkül, néhány standard módszerrel. A jellemző lemmák tekintetében a J48-as fa teljesített a legjobban, a Köszönet szerkezeti egység a kizárólagos lemmatizálással hozta a legjobb eredményt, 86%-ot,[6] a többi dimenzió csak rontott ezen. A többi szerkezeti egység tekintetében a lemmatizálás azonban önmagában kiemelkedő eredményt nem hozott.

[6] A tanulmányban szereplő százalékos értékek az F-mértékre vonatkoznak.

### 3.2. Lexikai és szemantikai jellemzők

A következő kísérletünk az volt, hogy a korpuszunkban az absztrakt címkével ellátott szavakat lecseréltük az absztrakt címkének a nevére, tehát például az OTP-t SZERVEZET-re cseréltük minden előfordulás esetén. Ezt a változatot tehát lemmák, illetve egyes lemmák helyett absztrakt címkék alkották. Ez természetesen némileg csökkentette a teljes korpusz lemmáinak számát. Ez a kísérlet már hozott némi javulást, ezzel értük el a legjobb eredményeket az Elvárás, a Lezárás, valamint a Cél tekintetében. Az Elvárás és a Lezárás felismerését az SMO algoritmus javította, míg a Cél felismerését a Naive Bayes, l. 1. és 2. táblázat[7].

1. táblázat. Lexikai és szemantikai jellemzőkre vonatkozó kísérletek eredményei Naive Bayes algoritmussal

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.839 | 0.011 | 0.839 | 0.839 | 0.839 | 0.97 | Bemutatkozas |
| **0.571** | **0.007** | **0.727** | **0.571** | **0.64** | **0.931** | **Cel** |
| 0.789 | 0.007 | 0.833 | 0.789 | 0.811 | 0.983 | Csatolmany |
| 0.47 | 0.079 | 0.492 | 0.47 | 0.481 | 0.775 | Egyebkorulmenyek |
| 0.556 | 0.015 | 0.417 | 0.556 | 0.476 | 0.894 | Elismeres |
| 0.794 | 0.078 | 0.726 | 0.794 | 0.759 | 0.891 | Elvaras |
| 0 | 0 | 0 | 0 | 0 | 0.94 | Javaslat |
| 0.72 | 0.02 | 0.667 | 0.72 | 0.692 | 0.977 | Koszonet |
| 0.632 | 0.051 | 0.343 | 0.632 | 0.444 | 0.95 | Lezaras |
| 0.919 | 0.008 | 0.958 | 0.919 | 0.938 | 0.985 | Megszolitas |
| 0.621 | 0.07 | 0.742 | 0.621 | 0.676 | 0.914 | Problema |
| 0.705 | 0.05 | 0.719 | 0.705 | 0.708 | 0.913 | Weighted Avg. |

### 3.3. Lexikai, szemantikai és grammatikai jellemzők

A lemmák és az absztrakt címkék mellett vizsgáltuk még, hogy az igeidő (jelen vs. múlt) milyen szerepet játszik a szerkezeti egységek felismerésében. Ennek előzménye az volt, hogy a korpusz tanulmányozása során feltűnt, hogy azt a problémáját, amelyre segítséget vár, a levélíró múlt időben ismerteti, míg az Egyéb körülményekben előforduló szánalmas életkörülményekre való hivatkozás gyakran jelen időben történik. Ennek oka feltehetően az, hogy a megoldásra váró probléma általában egy múltban történt eseménynek vagy események sorozatának a közvetlen következménye, és a levélíró ezt az eseményt vagy események sorozatát részletezi a problémája ismertetésekor, míg a levélíró szánalomra méltó életkörülményei esetében inkább azt hangsúlyozza, hogy azok a jelenben is fennállnak, mintegy tetézik a bajt. Ennek megfelelően a magyarlanc által igéknek

[7] Az aláhúzás a 70% fölötti eredményeket jelöli, a vastag betűs kiemelés pedig az adott szerkezeti egység azonosításában elért legjobb eredményt az összes szempontot és eszközt figyelembe véve.

2. táblázat. Lexikai és szemantikai jellemzőkre vonatkozó kísérletek eredményei SMO algoritmussal

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.839 | 0.02 | 0.743 | 0.839 | 0.788 | 0.977 | Bemutatkozas |
| 0.5 | 0.018 | 0.467 | 0.5 | 0.483 | 0.814 | Cel |
| 0.789 | 0.009 | 0.789 | 0.789 | 0.789 | 0.929 | Csatolmany |
| 0.576 | 0.104 | 0.475 | 0.576 | 0.521 | 0.841 | Egyebkorulmenyek |
| 0.556 | 0.015 | 0.417 | 0.556 | 0.476 | 0.804 | Elismeres |
| **0.814** | **0.064** | **0.767** | **0.814** | **0.79** | **0.938** | **Elvaras** |
| 0 | 0 | 0 | 0 | 0 | 0.471 | Javaslat |
| 0.76 | 0.013 | 0.76 | 0.76 | 0.76 | 0.963 | Koszonet |
| **0.684** | **0.015** | **0.65** | **0.684** | **0.667** | **0.975** | **Lezaras** |
| 0.919 | 0.023 | 0.883 | 0.919 | 0.901 | 0.982 | Megszolitas |
| 0.586 | 0.048 | 0.8 | 0.586 | 0.677 | 0.884 | Problema |
| 0.718 | 0.047 | 0.729 | 0.718 | 0.718 | 0.916 | Weighted Avg. |

ítélt tokenek mellé felvettük egy külön dimenzióként, hogy azok múlt vagy jelen idejűek. Az eredmények azonban nem feleltek meg a várakozásoknak, egyik algoritmussal sem hoztak jelentős javulást a puszta lemmatizáláshoz képest. További paraméter volt a szerkezeti egységek azonosításában az igemód, hiszen az elvárásait a levélíró az esetek döntő többségében expliciten és felszólító módban fogalmazza meg, míg a problémáját jellemzően kijelentő módban. Az absztrakt címkék és az igemód a Naive Bayes algoritmussal az Elismerés felismerését javította, ám ebben az esetben is igen gyenge eredményt értünk csak el, 52%-os pontosságot. Az Elismerés szerkezeti egység felismerésének tekintetében azonban ez volt a legjobb eredményünk.

Mind az igeidőnél, mind az igemódnál elmondható, hogy bár a döntési fákban előkelően közel kerültek a gyökérhez, tehát szignifikánsak, a globális pontosságon mégsem tudtak érdemben javítani.

### 3.4. Lexikai és szerkezeti jellemzők

Ebben az esetben a lemmatizálást kiegészítettük a szerkezeti jellemzők tesztelésével is, úgymint a szerkezeti egységek levélen belüli abszolút elhelyezkedése, valamint a szerkezeti egységek egymáshoz viszonyított elhelyezkedése. Ehhez a következő dimenziókkal egészítettük ki a szerkezeti egységeket:

- CU_START_Q1..4 -> A levél melyik negyedében kezdődik a példány
- CU_START_D1..10 -> Melyik tizedben kezdődik a példány
- CU_LENGHT_Q1..4 -> A 100 százalék hány negyedét teszi ki a szerkezeti egység hossza
- CU_LENGHT_Q1..10 -> A 100 százalék hány tizedét teszi ki a szerkezeti egység hossza
- CU_PRECEDING_Megszolitas, Bemutatkozas, stb: a példány előtt ilyen szerkezeti egységek találhatók

– CU_FOLLOWING_Megszolitas, Bemutatkozas, stb: a példány után ilyen szerkezeti egységek találhatók

Itt több lépésben végeztük kísérleteinket. Vizsgáltuk egyrészt (1) csak a lemmákat és a szerkezeti egység abszolút elhelyezkedését, (2) a lemmákat, a szerkezeti egység abszolút elhelyezkedését és hosszát, valamint (3) a lemmákat és a szerkezeti egység abszolút és relatív elhelyezkedését. Ezek a kísérletek már hoztak jelentős javulást a felismerésben. Az (1) esetben a Naive Bayes algoritmus hozta a Bemutatkozás és a Megszólítás szerkezeti egységek tekintetében a legjobb eredményeket, rendre 89%-os és 96%-os felismerési pontosságot, l. 3. táblázat. Az SMO algoritmus pedig a Csatolmány, valamint az Elvárás szerkezeti egységek felismerési pontosságát maximalizálta, a Csatolmány esetében 89%-ra, míg az Elvárás esetében 79%-ra.

3. táblázat. Lexikai és szerkezeti jellemzőkre vonatkozó kísérletek eredményei Naive Bayes algoritmussal

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| **0.867** | **0.005** | **0.929** | **0.867** | **0.897** | **0.996** | **Bemutatkozas** |
| 0.722 | 0.04 | 0.419 | 0.722 | 0.531 | 0.965 | Cel |
| 0.947 | 0.02 | 0.667 | 0.947 | 0.783 | 0.991 | Csatolmany |
| 0.491 | 0.072 | 0.483 | 0.491 | 0.487 | 0.827 | Egyebkorulmenyek |
| 0.182 | 0.009 | 0.333 | 0.182 | 0.235 | 0.945 | Elismeres |
| 0.684 | 0.053 | 0.765 | 0.684 | 0.722 | 0.913 | Elvaras |
| 0 | 0 | 0 | 0 | 0 | 0.081 | Javaslat |
| 0.846 | 0.018 | 0.733 | 0.846 | 0.786 | 0.992 | Koszonet |
| 0.625 | 0.033 | 0.4 | 0.625 | 0.488 | 0.982 | Lezaras |
| **0.938** | **0.003** | **0.987** | **0.938** | **0.962** | **0.994** | **Megszolitas** |
| 0.686 | 0.068 | 0.771 | 0.686 | 0.726 | 0.915 | Problema |
| 0.722 | 0.042 | 0.737 | 0.722 | 0.724 | 0.933 | Weighted Avg. |

A (2) kísérlet az egyeshez hasonlóan javította a Megszólítás szerkezeti egység felismerését (szintén 96%) a Naive Bayes algoritmus esetében, a Bemutatkozás tekintetében azonban gyengébben szerepelt.

A (3) esetben az Egyéb körülmények és a Probléma szerkezeti egységek tekintetében értünk el maximum pontosságot. Az Egyéb körülmények esetében ez 55%-os pontosságot jelent, míg a Probléma esetében jobb az eredmény, 78%.

### 3.5. Összevont kategóriák

Érdekesnek találtuk megvizsgálni azt is, hogy a tanítás szempontjából melyek azok a szerkezeti egységek amelyeket könnyen azonosít a gép és melyek azok, amelyeket nehezen, és hogyan lehetne összevonni az egyes szerkezeti egységeket, ha nagy pontosságú, de kisebb felbontású kategorizálásra lenne szükség. A konfúziós mátrix megvizsgálásával az alábbi három csoport esik közel egymáshoz:

Bemutatkozás, Megszólítás, Cél, Csatolmány (BeMegCelCsat) Probléma, Javaslat, Elvárás, Elismerés, Egyéb körülmények (ProElisEgyebElvarJav) Köszönet, Lezárás (KoLe). Ezen csoportokra is lefuttattuk a kategorizáló eljárásokat a lexikai és a szemantikai jellemzők (azaz a lemmák és az absztrakt címkék) figyelembevételével, és nem meglepő módon 90% körüli eredményeket kaptunk. A legjobb eredményt az SMO algoritmussal értük el, minden csoport esetén 85% fölötti pontossággal, l. 4. táblázat.

4. táblázat. Az összevont kategóriák tesztelésének eredményei SMO algoritmussal

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| **0.887** | **0.056** | **0.881** | **0.887** | **0.884** | **0.925** | **BeMegCelCsat** |
| **0.826** | **0.012** | **0.884** | **0.826** | **0.854** | **0.948** | **KoLe** |
| **0.942** | **0.092** | **0.935** | **0.942** | **0.938** | **0.93** | **ProElisEgyebElvarJav** |
| 0.913 | 0.073 | 0.913 | 0.913 | 0.913 | 0.93 | Weighted Avg. |

### 3.6. Az eredmények összegzése

Az egyes szerkezeti egységek felismerésében tehát 90% fölötti pontosságot csak a Megszólítás esetében sikerült elérni, amelynek a megformálása a legkonvencionálisabb módon történik, így ez az eredmény nem is meglepő. A konvencionális formák a mentális lexikonban egy egységként tárolódnak, előre megkomponáltak, így rutinszerűen működtethetők, különösebb kreativitást nem igényelnek a beszélőtől ([13]: 23). 80% fölötti pontosságot értünk el a Bemutatkozás, a Köszönet, valamint a Csatolmány szerkezeti egységek felismerésében, amely szerkezeti egységek (vagy legalábbis egy részük) megformálása szintén erősen konvencionális. 70% fölötti a pontossága a Problémának és az Elvárásnak, 60% fölötti a Célnak és a Lezárásnak és a leggyengébb, azaz 50% fölötti a pontossága az Egyéb körülmények és az Elismerés szerkezeti egységeknek. Könnyen belátható, hogy ezeknek a megformálása már sokkal szabadabb, az egy Lezárást kivéve, amelynek a definíciója feltehetően pontosításra szorul. A Javaslat szerkezeti egységre nem volt példa a tesztadatok között.

Az egyes szerkezeti egységekre lebontva az eredményeket a következőképpen összegezhetjük. A Bemutatkozás szerkezeti egység felismeréséhez a legnagyobb mértékben a lemmatizálás és a szerkezeti egységek abszolút elhelyezkedése járult hozzá, 14%-os javulást hozva az egyedüli lemmatizáláshoz képest. A Cél felismeréséhez az absztrakt címkék megoszlása járult hozzá a legnagyobb mértékben, a lemmatizáláshoz képest szintén 14%-os javulást eredményezett. Az Egyéb körülmények szerkezeti egység felismerését a legnagyobb mértékben a szerkezeti egységek abszolút és relatív elhelyezkedése, valamint hosszúsága és a lemmatizálás javította 11%-kal. Az Elismerés felismerésében a legjobb eredményt az absztrakt címkék és az igemód vizsgálata hozta, a puszta lemmák figyelembevételéhez képeset 10%-os javulást eredményezve, az Elvárás felismerését pedig

az absztrakt címkék vizsgálata, valamint a szerkezeti egységek abszolút és relatív elhelyezkedése a lemmatizálással együtt azonos mértékben javította a puszta lemmatizáláshoz képest, 9%-os javulást hozva. A Köszönet szerkezeti egység felismerésében önmagában a lemmatizálással kaptuk a legjobb eredményt, a Lezárás azonosításához a legnagyobb mértékben az absztrakt címkék és a lemmatizálás járult hozzá, az egyedüli lemmatizáláshoz képest 15%-kal javítva a pontosságot. A Megszólítás szerkezeti egység felismerésében a legjobb eredményt a lemmatizálással és a szerkezeti egységek abszolút elhelyezkedésével értük el, a puszta lemmatizáláshoz képest 13%-kal jobb eredményt értünk el. A Probléma szerkezeti egység felismerésében a legjobb eredményt a lemmatizálás, a szerkezeti egységek abszolút és relatív elhelyezkedése, valamint a hosszúsága hozta, 15%-os javulást a puszta lemmatizáláshoz képest. Végül a Csatolmány felismerésében a legjobb eredményeket az absztrakt címkék, valamint a lemmatizálás és a szerkezeti egységek abszolút és relatív elhelyezkedése hozta.

A szerkezeti egységek egyenkénti vizsgálati eredményeit jelentősen javították, amikor ezeket a konfúziós mátrixból kiolvasható szisztematikus tévesztések szerint három nagy csoportba vontuk össze. Ezek azonosításában a lemmákat és az absztrakt címkéket vettük figyelembe. Az azonosításban a legjobb eredményt az SMO algoritmussal értük el. Ebben az esetben viszont nem tudtuk elkülöníteni egymástól a Probléma és az Egyéb körülmények szerkezeti egységeket, amit viszont szerettünk volna.

Meglátásunk szerint növelné a találati pontosság hatékonyságát, ha a szerkezeti egységek felismerését két lépésben végeznénk. Első lépésben az összevont kategóriák felismerése történne, majd második lépésben ezeket a szerkezetiegység-csoportokat bontanánk tovább az egyes önálló szerkezeti egységekre. Külön érdemesnek tartanám megvizsgálni a következőket: ige-főnév eloszlást a KoLe csoportban, valamint igeidőt és igemódot vizsgálni a ProElisEgyebElvarJav csoportban, ezzel javítani a Probléma és Egyéb körülmények egymástól szétválasztását, valamint az Elvárás leválasztását. Emellett szükségesnek látszik kézi annotálással jelölni az egyes szerkezeti egységekre jellemző kifejezéseket, konvencionalizálódott szókapcsolatokat, ezzel segítve a szerkezeti egységek pontosabb gépi felismerését.

## 4. Összefoglalás

A tanulmány elsődleges célja annak bemutatása volt, hogy milyen módon nyerhetők ki a tematikus kontextusra vonatkozó információk a panaszlevelekből felszíni jellemzők alapján. A panaszleveleket tehát szerkezeti egységekre bontottuk, amelyek a bennük előforduló nyelvi információ tematikus kontextusaként szolgáltak. Ezután azt teszteltük, hogy ezeknek a szerkezeti egységeknek a gépi felismerését milyen felszíni jellemzők segítik elő. A szokásos lexikai tulajdonságok mellett vizsgáltuk még a kifejezések szemantikai és grammatikai tulajdonságait is, valamint a szerkezeti egységek egymáshoz képesti és abszolút elhelyezkedését. A jellemző lemmák vizsgálatához képest az extra dimenziók vizsgálata átlag 11%-os

javulását eredményezett, amely az esetek kb. felében szignifikánsnak tekinthető. A legjobb eredményt az összevont szerkezetiegység-csoportokkal értük el.

Célunk volt még, hogy felhívjuk a figyelmet arra, hogy szabadon alkotott szövegek esetében érdemes lehet a szerkezeti jellemzők figyelembevétele, tekintve, hogy a szövegen belüli tematikus kontextus felismerése pontosabb tartalmi kivonatolást tesz lehetővé.

## Hivatkozások


1. Sperber, D., Wilson, D.: Relevance: Communication and Congition. Blackwell, Oxford (1986/1995)
2. Rott, H. In: Words in Context: Fregean Elucidations. Volume 23. (2000) 621–641
3. Bunt, H., Black, B. In: The ABC of computational linguistics. Benjamins, Amsterdam (2000) 1–46
4. Bibok, K., Németh T., E. In: Lexikai és kontextuális információk interakciója a megnyilatkozásjelentés megalkotása során. Blackwell, Szeged (2002) 335–368
5. Carston, R. In: Relevance Theory and the Saying/Implicating Distinction. MIT Press, Cambridge MA (2004) 633–656
6. Németh T., E., Bibok, K.: Interaction between Grammar and Pragmatics: The Case of Implicit Arguments, Implicit Predicates and Co-composition in Hungarian. Journal of Pragmatics **4** (2010) 501–524
7. Tátrai, S.: A kontextus fogalmáról. **128**(4) (2004) 479–494
8. Zsibrita, J., Nagy, I., Farkas, R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: Magyar Számítógépes Nyelvészeti Konferencia. (2009)
9. Quinlan, J.R.: C4.5: Programs for machine learning (1992)
10. John, G., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann (1995) 338–345
11. Platt, J.C.: Fast training of support vector machines using sequential minimum optimization. In: Smola (Eds.), Advances in Kernel Methods-Support Vector Learning, MIT Press (1998) 185–208
12. Holmes, G., Donkin, A., Witten, I.: Weka: A machine learning workbench. In: Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia (1994)
13. Szili, K.: A kérés pragmatikája a magyar nyelvben. **126**(1) (2002) 12–30