Szintaktikai elemzés

Simon Eszter

2023. április 17.

A szintaktikai elemzés az a feladat, amelynek során szintaktikai szerkezetet rendelünk egy mondathoz. A szintaktikai elemzést nem csak úgy *l'art pour l'art* végezzük – magasabb szintű nyelvfeldolgozó alkalmazásokhoz szükséges a kimenete. Például a szemantikai elemzéshez, a kérdésmegválaszoláshoz vagy az információkinyeréshez kifejezetten hasznos, ha azonosítani tudjuk a mondat egyes szereplőit.

Háromféle szintaktikai elemzésről lesz szó: konstituenselemzés, dependenciaelemzés és sekély szintaktikai elemzés.

1. Konstituenselemzés

1.1. Konstituens

A konstituens (vagy más néven összetevő) szavak csoportja, amely egy egységként viselkedik. Például van a főnévi frázis (noun phrase, NP), amely egy főnév köré szerveződő egységet alkot, például:

(1) Micimackó

(2) a kutyák

(3) én

(4) azok a kutyák, amelyeket a szomszéd vásárolt a járvány kitörése előtt egy nappal a kisunokájának

Honnan tudjuk, hogy mely szavak alkotnak egy konstituenst? Az egyik jel, ha hasonló szintaktikai környezetben szerepelnek. Az 1–4. példákban szereplő NP-k például mind lehetnek egy mondat alanyai (a szóban forgó NP-ket *dőlt betűvel* jelzem):

(5) *Micimackó*, mint minden medve, szereti a mézet.

(6) *A kutyák* cukik.

(7) *Én* még nem kaptam el a vírust.

(8) Iszonyúan idegesítenek *azok a kutyák, amelyeket a szomszéd vásárolt a járvány kitörése előtt egy nappal a kisunokájának*.

Miközben ezek a komplett frázisok szerepelhetnek ugyanolyan pozícióban, a bennük szereplő egyes szavak önmagukban nem feltétlenül. Persze ez csak akkor érdekes, ha többszavas a frázis:

(9) **A* cukik.

(10) *Iszonyúan idegesítenek *vásárolt*.

Egy másik jele annak, hogy bizonyos szavak egy frázist alkotnak, hogy miközben a teljes frázis mozgatható a mondatban, az egyes szavak nem.

(11) A mézet szereti *Micimackó*, mint minden medve.

(12) Cukik *a kutyák*. DE: **A* cukik *kutyák*.

(13) A vírust még nem kaptam el *én*.

(14) *Azok a kutyák, amelyeket a szomszéd vásárolt a járvány kitörése előtt egy nappal a kisunokájának*, iszonyúan idegesítenek. DE: *Iszonyúan *azok* idegesítenek *a kutyák, amelyeket a szomszéd vásárolt a járvány kitörése előtt egy nappal a kisunokájának*.

1.2. Környezetfüggetlen nyelvtan

A nyelvek összetevős szerkezetének modellezésére legtöbbször használt formális rendszer a környezetfüggetlen nyelvtan (Context-Free Grammar, CFG) vagy más néven frázisstruktúra nyelvtan (Phrase Structure Grammar). Egy környezetfüggetlen nyelvtan áll egyrészt egy szavakat és szimbólumokat tartalmazó lexikonból, másrészt szabályok halmazából, amelyek azt szabják meg, hogy ezek a szimbólumok hogyan tudnak egybecsoportosulni. Ezeket hívják derivációnak, például:

(15) S → NP VP

(16) NP → Det N

(17) Det → a

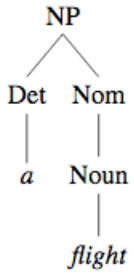
(18) N → kutya

A 15. és a 16. példában szereplő szimbólumokat nem-terminális szimbólumoknak, a 17. és a 18. példában szereplőket pedig terminális szimbólumoknak hívjuk. A CFG-ben egy szabály bal oldalán mindig egy nem-terminális szimbólum van, míg a jobb oldalán lehet egy vagy több terminális vagy nem-terminális. Vegyük észre, hogy azokban a szabályokban, amelyeknek a jobb oldalán egy terminális van, a bal oldalon levő nem-terminális a terminálisnak a szófaji címkéje.

A szabály jobb oldalán több nem-terminális elem is állhat, de a példákban végig Chomsky-féle normálalakban (Chomsky normal form, CNF) közlöm a szabályokat, ami szerint egy szabály jobb oldalán vagy két nem-terminális, vagy egy terminális lehet ($A \rightarrow B\ C$ vagy $A \rightarrow a$) [Chomsky, 1963]. Természetesen a Chomsky-féle normálalak és bármilyen más formában közölt CFG szabály átkonvertálható, például így: $A \rightarrow B\ C\ D$ helyett $A \rightarrow B\ X$ és $X \rightarrow C\ D$. A Chomsky-féle normálalak bináris elágazást valósít meg, ami kisebb nyelvtanokhoz vezet, ami komputációs szempontból kifejezetten hasznos.

Egy CFG-re két irányból is lehet tekinteni: egyrészt lehet egy eszköz arra, hogy generáljunk mondatokat, másrészt meg arra, hogy elemezzünk mondatokat, vagyis egy szerkezetet rendeljünk egy mondathoz. Ha generátorként tekintünk rá, akkor a szabályban található nyíl azt jelenti, hogy „írd újra a bal oldalon levő szimbólumot a jobb oldalon levő szimbólumokkal". Ezért szokták újraíró szabályoknak is hívni őket. Így lesz a fenti példaszabályok alapján az NP-ből *a kutya*.

A CFG által létrehozott szerkezetet általában faként szokták ábrázolni (igazából fordított fa, mert a gyökércsomópont van felül). A gyökércsomópont kitüntetett pozíció: start szimbólumnak hívják, és jellemzően S-sel jelölik. Egy ilyen fában az egyes csomópontok mindig dominálják az alattuk levőket. Egy kis példafa látható az 1. ábrán.



1. ábra. Egy kis példafa: az angol *a flight* főnévi frázishoz tartozó elemzési fa.

A főnévi frázisok mellett természetesen léteznek más frázisok is: igei frázis, posztpozíciós frázis, melléknévi frázis, határozói frázis:

(19) Iván süt nekem sütit (VP → NP V NP NP)

(20) a sütő mellett (PostP → NP Post)

(21) régi sütő (AP → A N)

(22) régen (AdvP → Adv)

A faábrázolás mellett szoktak zárójelezést is használni, ilyenkor a nyitó zárójel mellé alsó indexbe szokták írni a frázis típusát, ahogy az a 2. ábrán látható.

Azok a mondatok, amelyeket egy CFG szabályai segítségével elő tudunk állítani, grammatikus mondatok, amelyek pedig nem állíthatók elő a szabályok

$[_S\ [_{NP}\ [_{Pro}\ \text{I}]]\ [_{VP}\ [_V\ \text{prefer}]\ [_{NP}\ [_{Det}\ \text{a}]\ [_{Nom}\ [_N\ \text{morning}]\ [_{Nom}\ [_N\ \text{flight}]]]]]]$

2. ábra. Az *I prefer a morning flight* mondat elemzése zárójeles notációval.

segítségével, azok agrammatikus mondatok; vagyis a nyelv egyenlő a nyelvtan által generált összes lehetséges mondat halmazával.

1.3. Bizonyos nyelvi jelenségek kezelése a konstituenselemzésben

A CFG megszabja a terminális szimbólumok sorrendjét is, vagyis ha az a szabályunk, hogy $S \rightarrow NP\ VP$, akkor a főnévi frázisnak meg kell előznie az igét. De vannak olyan nyelvek, amelyeknek meglehetősen szabad a szórendjük, és az egyes – amúgy összetartozó – mondatrészek el tudnak mozogni egymástól. Ezeket hívják távoli függőségeknek, mint például a *Jánosnak tegnap láttam utoljára a kalapját* mondatban, ahol a *Jánosnak* és az *a kalapját* egy birtokos szerkezet összetartozó elemei. A frázisstruktúra nyelvtanokban ezt a jelenséget úgy kezelik, mintha az adott frázis az adott helyen keletkezett, majd onnan elmozgott volna, egy üres pozíciót hagyva maga után, ami össze van indexelve az új hellyel.

Minden frázisnak van feje, ami a centrális eleme az adott frázisnak. Ez a fej határozza meg a frázist: a főnévi frázis feje jellemzően egy főnév, egy igei frázis feje egy ige.

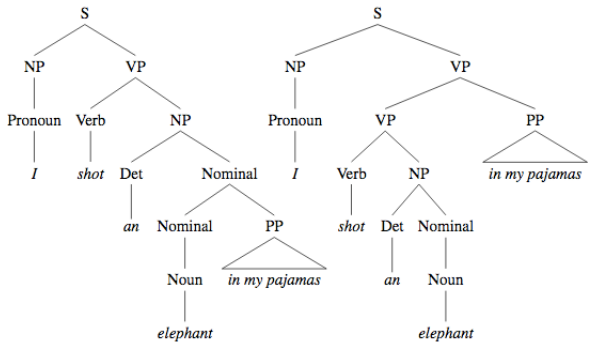
A fej grammatikai szempontból a legfontosabb elem a frázisban. Kiemelten fontos szerepe van az olyan konstituensalapú nyelvtanok esetében, mint a fejközpontú frázisstruktúra nyelvtan (Head-Driven Phrase Structure Grammar, HPSG) [Pollard and Sag, 1994] vagy a dependenciaelemzés, amelyről a 2. fejezetben lesz szó.

A többértelműség kérdése a szintaxis szintjén is problémát okoz. Szerkezeti többértelműségről beszélünk akkor, amikor egy nyelvtan egy mondathoz több helyes elemzést is tud társítani. Itt van például az alábbi idézet:

> „One morning I shot an elephant in my pajamas. How he got in my pajamas, I don't know." (Groucho Marx, *Animal Crackers*, 1930)

A szerkezeti többértelműségnek az egyik fajtája az, ami ebben az idézetben is megjelenik (attachment ambiguity): a többértelműség onnan ered, hogy az *in my pajamas* prepozíciós frázist több helyre is lehet csatolni az elemzési fában, ahogy az a 3. ábrán látható. Ha az elefánt van pizsamában, akkor az *elephant* mellé csatolódik, ha pedig a mondat alanya van pizsamában, akkor a VP-vel van egy szinten a fában.

Egy másik fajta szerkezeti többértelműség a koordinációs többértelműség, amikor ugyanolyan típusú frázisok kapcsolódnak össze egy *és*-sel, de nem tudjuk eldönteni, hogy hogyan zárójeleződnek. Például az *öreg férfiak és nők* frázis esetében két zárójelezési lehetőség van: *(öreg férfiak) és (nők)* vagy *(öreg (férfiak és nők))*.



3. ábra. Két elemzési fa egy többértelmű mondathoz.

Számos olyan szintaktikai többértelműség van, ami szemantikailag értelmetlen. A szintaktikai egyértelműsítés során az a feladat, hogy az összes lehetséges szintaktikai elemzés közül kiválasszuk a legjobbat. Ha egy elemző ehhez statisztikai, szemantikai és kontextuális információkat is fel tud használni, akkor hatékonyabb lesz.

1.4. Treebankek

A szintaktikai elemzés során tehát egy mondathoz egy elemzési fát illesztünk. Ebből az következik, hogy létre lehet hozni mondatoknak egy olyan gyűjteményét, amiben minden egyes mondathoz hozzá van rendelve egy elemzési fa. Ezeket a szintaktikailag annotált korpuszokat hívják treebankeknek. Ezek nagy része félig automatikusan készült, vagyis egy automatikus szintaktikai elemzővel először leelemezték az összegyűjtött mondatokat, majd kézzel ellenőrizték és javították.

A legismertebb treebankek egyike a Penn Treebank[1]. A 4. ábrán a Penn Treebankből látunk egy példamondatot, amiben az elmozgott mondatrészek indexekkel vannak jelölve, ahogy azt említettük az 1.3. fejezetben.

A mondatok egy treebankben implicite annak a nyelvnek a nyelvtanát adják, amit a korpusz reprezentál. Vagyis egy létező treebankből a szabályokat visszafejtve egy CFG-t kapunk.

1.5. Lexikalizált nyelvtanok

A fent bemutatott nyelvtanok esetében a fókusz a szabályokon van. A rengeteg szabály komputációs szempontból nem praktikus – egyes esetekben a lehetséges elemzések száma exponenciálisan nő. Ezen kívül ezek az összetevős nyelvtanok nehezen kezelik az olyan nyelvi jelenségeket, mint a távoli függőségek, az egyeztetés vagy az igék szubkategorizációs keretei. Ezért számos olyan nyelvtant fejlesztettek ki, amelyekben a szabályokról a fókusz a lexikonra tevődik át. Ezek

[1] `https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html`

```
( (S (`` ``)
    (S-TPC-2
      (NP-SBJ-1 (PRP We) )
      (VP (MD would)
        (VP (VB have)
          (S
            (NP-SBJ (-NONE- *-1) )
            (VP (TO to)
              (VP (VB wait)
                (SBAR-TMP (IN until)
                  (S
                    (NP-SBJ (PRP we) )
                    (VP (VBP have)
                      (VP (VBN collected)
                        (PP-CLR (IN on)
                          (NP (DT those)(NNS assets)))))))))))))
    (, ,) (`` ``)
    (NP-SBJ (PRP he) )
    (VP (VBD said)
      (S (-NONE- *T*-2) ))
    (. .) ))
```

4. ábra. Példamondat a Penn Treebankből. Az elmozgott mondatrészek indexekkel vannak jelölve.

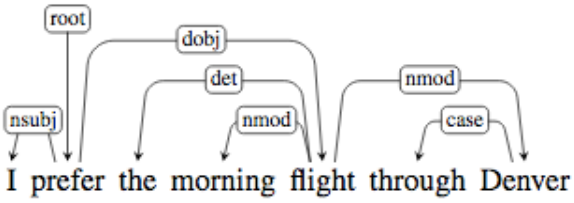
az úgynevezett lexikalizált nyelvtanok, amelyek többek közt abban különböznek egymástól és az eddig ismertetettektől, hogy mennyire lexikalizáltak.

Több lexikalista elmélet is született, amelyek legfontosabb jellemzője, hogy megszorításalapúak, vagyis nem egy sikeres deriváció, hanem egy kielégíthető követelményhalmaz a grammatikus nyelvi formák ismérve. A legfontosabb megszorításalapú elméletek a lexikai-funkcionális nyelvtan (Lexical-Functional Grammar, LFG) [Kaplan and Bresnan, 1982], a generalizált frázisstruktúra nyelvtan (Generalized Phrase Structure Grammar, GPSG) [Gazdar et al., 1985], majd a már említett utódja, a HPSG, a kategoriális nyelvtanok (Categorial Grammar, CG) és a konstrukciós nyelvtan [Goldberg, 1995]. Egyes elméletek még ennél is tovább mentek, és radikális lexikalizmust hirdettek, ami szerint ha elég gazdag a lexikon, akkor pusztán függvényalkalmazás és unifikáció segítségével is összeépíthető a mondat, nincs szükség egyéb szintaktikai műveletre, továbbá a frázisstruktúra-fa sem fog többé valódi információt hordozni. Vagyis a nyelvi jelenségek kezelése megmarad, csak ezek megragadása az egyes elméletek esetében máshol történik: ha nem szabályokkal, akkor a lexikonban.

2. Dependenciaelemzés

A szintaktikai elemzés egy másik fajtája a dependenciaelemzés vagy más néven függőségi elemzés. Ebben a formalizmusban nem az összetevők és a szabályok

játsszák a központi szerepet – a mondat szintaktikai szerkezetét a szavak és a köztük levő irányított bináris relációk adják, ahogy az az 5. ábrán látható. A szavak közötti relációk a mondat fölé húzott irányított, címkézett élekkel vannak ábrázolva, amelyek a szerkezet fejétől haladnak a tőle függő szavak felé. Az éleken levő címkék grammatikai relációk egy előre meghatározott halmazából kerülnek ki.



5. ábra. Az *I prefer the morning flight through Denver* mondat dependenciaelemzése.

A dependenciaelemzés előnye a konstituenselemzéssel szemben, hogy a távoli függőségeket is jól tudja kezelni, ami főleg a gazdag morfológiájú és szabad szórendű nyelvek esetében hasznos. A másik nagy előnye, hogy az egyes mondatrészek által betöltött grammatikai funkciókat, így például, hogy mi a mondat alanya és tárgya, is megtudjuk belőle. Ezek az információk pedig már a szemantikai elemzés felé nyitnak utat.

A jelenleg legismertebb és leginkább használt dependenciaelemzési keretet a Universal Dependencies[2] (UD) nyújtja. Ez egy nemzetközi projekt, amelynek keretében univerzális formalizmust és címkekészletet fejlesztettek ki elvileg az összes nyelvre. Ebben a formában treebankek is elérhetőek jelenleg több mint 130 nyelvre.

Egy függőségi mondatszerkezet formálisan: $G = (V, A)$, ahol a V olyan csúcsok halmaza, amelyek pontosan megfelelnek a mondat szavainak, és az A azon élek halmaza, amelyek a V-k közötti relációkat ragadják meg. Egy függőségi fa egy irányított gráf, amely megfelel az alábbi kritériumoknak:

1. Van egy darab speciális gyökércsomópont, amibe nem megy bele él.
2. A gyökércsomópont kivételével minden csomópontnak egy bejövő éle van.
3. Létezik egy egyedi útvonal a gyökércsomóponttól minden egyes csomópontig.

A treebankekben szereplő függőségi fák projektívek, ami azt jelenti, hogy nincsenek bennük keresztező élek. A függőségi fák általában környezetfüggetlen nyelvtanokból szoktak generálva lenni, ezért garantáltan projektívek.

A dependenciaelemzők kiértékelése során az automatikusan előállított elemzési fákat hasonlítják össze a gold standard treebankben szereplő fákkal. Viszont

[2]https://universaldependencies.org/

nem teljes mondatok elemzését vetik össze, mert akkor szinte soha nem kapnának egyezést, túl szigorú lenne a metrika, és ezáltal nem kapnának igazi képet a rendszer valódi teljesítményéről. Ehelyett két metrikát alkalmaznak: az egyik a *labeled attachment score* (LAS), a másik az *unlabeled attachment score* (UAS). Az első mérték azt számolja, hogy hány él van jól behúzva és egyszersmind jól felcímkézve, a másik pedig csak azt, hogy hány él van behúzva.

3. Sekély szintaktikai elemzés

Vannak olyan nyelvfeldolgozási feladatok, amelyek nem igényelnek teljes mondatelemzést, megelégszenek a sekély szintaktikai elemzéssel is. Ez utóbbi során annyi történik, hogy a mondat egyes összetevőit azonosítja az elemző, de arra vonatkozóan, hogy azok milyen viszonyban vannak egymással, nem mond semmit. Más néven *chunking*-nak (vagy chunkolásnak) is hívják ezt a feladatot, mivel a mondatot alkotó *chunk*-ok, vagyis frázisok megtalálása a cél. A chunkolás során a teljes mondatot lefedő, lapos, nem átfedő összetevőket azonosítjuk (23. példa). Van olyan eset is, amikor csak a mondatban levő főnévi frázisokat keressük, ezt hívjuk NP chunkolásnak. Ennek is két további fajtája van: az egyik esetében a legkisebb NP-ket (base NP) keressük (24. példa), amelyek nem tartalmaznak másik NP-t; a másik esetében a maximális NP-ket keressük, amelyek nem részei egy náluk nagyobb NP-nek (25. példa).

(23) $[_{NP}$ The morning flight] $[_{PP}$ from] $[_{NP}$ Denver] $[_{VP}$ has arrived].

(24) $[_{NP}$ The morning flight] from $[_{NP}$ Denver] has arrived.

(25) $[_{NP}$ The morning flight from Denver] has arrived.

A chunking során nem egy hierarchikus struktúrát tárunk fel, hanem valójában szegmentumok egy sorozatát keressük. Vagyis ez is felfogható szekvenciális címkézési feladatként, és ugyanúgy kezelhető, mint más hasonló típusú feladatok, mint például a tulajdonnév-felismerés. A chunkoláshoz használt state-of-the-art szekvenciális taggerek felügyelt gépi tanuláson alapulnak. A tanuló és a kiértékelő adathalmazban használt szekvenciákat az `IOB` prefixek valamelyikével szokták ellátni, ahol az `I` a szekvencia belsejét, az `O` a szekvencián kívüli elemeket, és a `B` a szekvencia kezdő elemét jelöli. A kiértékelés ugyanúgy zajlik, ahogy a tulajdonnév-felismerésnél: a pontosság, a fedés és az F-mérték kiszámolásával.

4. Magyar nyelvű elemzők és erőforrások

A legnagyobb magyar nyelvű gold standard treebank a Szeged Treebank[3], amely 82.000 mondatban 1,2 millió szót tartalmaz. A Szeged Treebank 2.0 – a mondatokra és tokenekre bontás, valamint az egyértelműsített morfológiai annotáció

[3] `https://rgai.inf.u-szeged.hu/node/113`

mellett – kézzel készített mély összetevős elemzést tartalmaz minden mondatra. Ezen a korpuszon lett tanítva a magyarlánc[4] és az e-magyar[5] összetevős elemzője, a magyarra adaptált Berkeley parser[6].

A Szeged Treebank szöveganyaga át lett konvertálva függőségi fákká is – így jött létre a Szeged Dependency Treebank. Ez szolgál gold standard tanítóanyagként a magyarláncba és az e-magyarba beépített Bohnet parsernek, ami a jelenleg elérhető legjobb teljesítményű dependenciaelemző magyarra (UAS: 93,22%, LAS: 91,42%) [Zsibrita et al., 2013].

A Universal Dependencies projekt keretei között elkészült egy kisebb magyar treebank is[7]. Ez a UD 2-es verziójának formalizmusát és címkekészletét követi, de csak 42.000 tokent (1800 mondatot) tartalmaz. Minden neurális hálós gépi tanuláson alapuló nyelvfüggetlen elemzőlánc (pl. HuSpaCy[8], UDPipe[9]) ezt a korpuszt használja tanítókorpuszként, aminek kis mérete miatt ezeknek az elemzőknek a teljesítménye nem annyira magas.

A HunTag3[10] szekvenciaelemző használható chunkolásra és tulajdonnév-felismerésre is, és ez lett beépítve az `e-magyar` szövegfeldolgozó eszközláncba is. Ennek eredeti verziója `hunchunk` néven futott, és a magyar nyelvű minimális NP-k felismerésében 95,48%-os, a maximális NP-k esetében pedig 89,11%-os F-mértéket produkált [Recski and Varga, 2010]. Szintén NP chunkolást valósít meg az `e-magyar`-ba beépített `emBERT` modul [Nemeskey, 2020], amely a BERT transformer encoderen alapul[11], és 95% felett teljesít mindkét NP chunkolási feladaton.

Hivatkozások

[Chomsky, 1963] Chomsky, N. (1963). Formal properties of grammars. In Luce, R. D., Bush, R. R., and Galanter, E., editors, *Handbook of Mathematical Psychology*, page 323–418. John Wiley and Sons, Inc.

[Gazdar et al., 1985] Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985). *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge.

[Goldberg, 1995] Goldberg, A. E. (1995). *Constructions*. The University of Chicago Press, Chicago and London.

[Kaplan and Bresnan, 1982] Kaplan, R. and Bresnan, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, page 173–281. The MIT Press.

[4] `https://rgai.inf.u-szeged.hu/node/100`
[5] `http://e-magyar.hu/hu/`
[6] `https://github.com/slavpetrov/berkeleyparser`
[7] `https://universaldependencies.org/treebanks/hu_szeged/index.html`
[8] `https://github.com/huspacy/huspacy`
[9] `http://ufal.mff.cuni.cz/udpipe`
[10] `https://github.com/nytud/HunTag3`
[11] `https://en.wikipedia.org/wiki/BERT_(language_model)`

[Nemeskey, 2020] Nemeskey, D. M. (2020). Egy `emBERT` próbáló feladat. In Berend, G., Gosztolya, G., and Vincze, V., editors, *XVI. Magyar Számítógépes Nyelvészeti Konferencia*, pages 411–420, 2020.

[Pollard and Sag, 1994] Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

[Recski and Varga, 2010] Recski, G. and Varga, D. (2010). A Hungarian NP Chunker. *The Odd Yearboook*, 8:87–93.

[Zsibrita et al., 2013] Zsibrita, J., Farkas, R., and Vincze, V. (2013). magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *International Conference on Recent Advances in Natural Language Processing*, pages 763–771, Shoumen, Bulgária. INCOMA Ltd.