Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Automatizálási és Alkalmazott Informatikai Tanszék

Morfológia a kontextuális nyelvi modellek korában

Tézisfüzet

Ács Judit

Témavezető:
Kornai András, D.Sc.

Budapest, 2025.

Bevezetés

A természetes nyelvfeldolgozás (NLP) területe az elmúlt 15 évben átvette a mélytanulási módszereket. Manapság a legtöbb NLP feladatban a legkorszerűbb megoldás valamilyen neurális modell, gyakran egy előre betanított nyelvi modell finomhangolt változata. E modellek hatékonyságát különböző angol nyelvű benchmark adatbázisokon, valamint egyre inkább más egynyelvű és többnyelvű adatbázisokon is bizonyítják.

Ebben a disszertációban a mélytanulási modellek alkalmazását vizsgálom alacsony szintű feladatokra, különösen morfoszintaktikai feladatokra több nyelven.

A disszertáció első része (1. és 2. tézis) a mélytanulási modellek alkalmazását tárgyalja klasszikus morfoszintaktikai feladatokra, mint például a morfológiai inflexió és elemzés több tucat nyelven, különös tekintettel a magyar nyelvre.

A disszertáció második része (3–5. tézis) az előre betanított nyelvi modellekkel, főként a BERT (Devlin et al., 2019) család modelljeivel, foglalkozik. Néhány kísérletet bemutatok a GPT-4o és a GPT-4o-mini modellekkel is. Ezek a modellek kiváló teljesítményt mutatnak különböző angol nyelvű és néhány sok erőforrással rendelkező nyelvű feladatokon. Ugyanakkor közepes és kevés erőforrással rendelkező nyelvek esetében a kiértékelésük hiányos. Bemutatok egy módszertanot a morfoszintaktikai benchmarkok generálására tetszőleges nyelveken, és részletesen elemzek több BERT modellt. Az elemzés fő eszköze a szondázó módszertan (Belinkov, 2021).

Kutatási módszertan

A disszertációm fő kutatási kérdése az, hogy a mélytanulási modellek mennyire jól kezelik a morfológiát különböző nyelveken. Ezt a kérdést két szempontból vizsgálom. Az első (I. rész, 1. és 2. tézis) kis méretű, kifejezetten alacsony szintű feladatokra tanított mélytanulási modelleket vizsgál. A második (II. rész, 3–5. tézis) közepes méretű, előre betanított nyelvi modelleket használ enkóderként, és ezek morfoszintaktikai tartalmát elemzem.

A disszertációm I. részében kézzel összeállított mélytanulási modelleket

alkalmazok, amelyeket end-to-end tanítással tanítok fel. A modelleket előtanítás nélkül tanítom kis- és közepes mennyiségű annotált adaton. A tanításhoz a szakirodalomban szokásos eljárásokat alkalmaztam, kivéve, ha másként nem jelzem.

A kutatás fő módszere a II. részben a szondázás (probing). A szondázást mint modellvizsgálati eszközt már számos területen alkalmazták, beleértve a morfológiát is, de nem olyan mélységben, mint ebben a disszertációban. A szondázás egyszerű és intuitív módja a modell belső tartalmának feltérképezésére, bár kritikusai néhány hibájára is rámutattak (Belinkov, 2021). Az ablációs vizsgálataim során kimutattam, hogy ezek a problémák csak korlátozott hatással vannak a kutatásaimra, és az eredményeim függetlenek a szondázás típusától és paramétereitől.

Újszerű perturbációkat vezetek be, amelyek eltávolítanak bizonyos információkat a mondatból, és a szondák újratanításával meg tudom állapítani, hogy ezek az információforrások szükségesek voltak-e. Ezt az elemzést játékelméletből átvett Shapley-értékekkel bővítem ki.

Hozzájárulásaimat az alábbi öt tézis foglalja össze.

1. tézis

Megmutattam, hogy az enkóder-dekóder (más néven sequence-to-sequence vagy seq2seq) modellek jól alkalmazhatók morfológiai inflexióra és generálásra. Ez igaz mind a szó-, mind a mondatszintű feladatokra több nyelven.

Altézisek:

1.1 Készítettem egy ezüst sztenderd magyar nyelvű adathalmazt morfológiai inflexió és elemzés céljából, jó minőségű szabályalapú elemző segítségével.

1.2 Implementáltam két típusú seq2seq vagy enkóder-dekóder modellt figyelemmechanizmussal: LSTM puha figyelemmel és LSTM szigorú figyelemmel.

1.3 Betanítottam és kiértékeltem a modelleket a magyar adathalmazon.

1.4 Két új típusú enkóder-dekóder modellt fejlesztettem ki morfológiai inflexiós feladatokra. Az első modell egy két enkóderes-egy dekóderes architektúra, amely típusszintű ragozásra szolgál. A második modell egy összetett, több enkóderből álló modell, amely mondatszintű ragozást végez. Ezekkel a modellekkel egyéni csapatként részt vettem a CoNLL-SIGMORPHON 2018 versenyen. Az 1. feladat típusszintű ragozás volt több mint 100 nyelven, a 2. feladat pedig mondatbeli (környezetfüggő) ragozás 7 nyelven. A két feladatban 3. és 2. helyezést értem el többenkóderes és szimpla dekóderes modelljeimmel.

A tézishez kapcsolódó eredményeket a Ács (2018)-ban publikáltam. A tézis minden eredménye a saját munkám.

2. tézis

Adaptáltam egy differenciálható véges állapotú súlyozott automata RNN-t sequence-to-sequence feladatokra és megmutattam, hogy a differenciálható neurális mintázatfelismerés több nyelven képes morfoszintaktikai minták kinyerésére, ha morfológiai inflexió és elemzés céljából enkóderként használják.

Altézisek:

2.1 Újraimplementáltam a Soft Patterns vagy SoPa neurális modellt (Schwartz et al., 2018), amely egy korlátozott és differenciálható véges automatákon alapuló modell, eredetileg szövegosztályozásra. Hozzáadtam egy LSTM dekódert, és SoPa-t enkóder-dekóder modellként használtam. A modell az enkóder oldalon mintázatfelismerést alkalmaz.

2.2 A modellt szótípus-szintű morfológiai elemzésre és ragozásra alkalmaztam 12 tipológiailag vegyes nyelven.

2.3 Kinyertem mintázatokat (karaktersorozatokat) a betanított SoPa modellekből, és manuálisan vizsgáltam azok nyelvészeti relevanciáját.

2.4 Bevezettem egy modellhasonlósági mérőszámot két SoPa modell összehasonlítására. Ez lehetővé teszi olyan feladatok összehasonlítását, ame-

lyek ugyanazt a bemeneti adatot használják. Minél nagyobb a hasonlóság két feladat között, annál valószínűbb, hogy ugyanazokat a mintázatokat használják.

Ezeket az eredményeket Ács and Kornai (2020) publikálta. A tanulmány elnyerte a legjobb cikk díját a 2020-as Magyar Számítógépes Nyelvészeti Konferencián. Az implementáció teljes mértékben az én munkám.

3. tézis

Kifejlesztettem egy új módszertant morfoszintaktikai szondázásra. A módszerem CoNLL-U formátumú adatokot használ, amelyek széles körben elérhetők az Universal Dependencies Treebankben, így a módszertanom sok nyelvre és morfoszintaktikai címkére alkalmazható. A szondázási módszerem segítségével kimutattam, hogy az előre betanított nyelvi modellek, amelyeket annotálatlan szövegen tanítottak, elsajátítják a morfológiát. A nyelvmodellek reprezentációi megtartják a morfoszintaktikai információkat, amit megmutattam tipológiailag változatos nyelveken és többféle feladaton keresztül.

Altézisek:

3.1 Létrehoztam a legnagyobb többnyelvű morfoszintaktikai szondázó adathalmazt, amely 247 feladatot tartalmaz 42 nyelven 10 nyelvcsaládból. Egy szondázó mintát egy mondat-token-morfoszintaktikai címke hármasként definiáltam. A minták generálásához a Universal Dependencies Treebank-et (Nivre et al., 2018) használtam.

3.2 Kiértékeltem 6 többnyelvű PLM-et, részletesen elemeztem az mBERTet és az XLM-RoBERTát, valamint összehasonlítottam őket különböző baseline modellekkel és megmutattam, hogy a PLM-ek valóban tanulnak morfológiát.

3.3 Kiértékeltem a GPT-4o és GPT-4o-mini modelleket a szondázó adathalmaz tesztkészletén, promptolás segítségével.

3.4 Megvizsgáltam a PLM-ek tokenizáló algoritmusát, és megmutattam, hogy bár a többnyelvű modellek több mint 100 nyelvet támogatnak, a tokenizálás jobban működik a latin betűs nyelveken, különösen az angol nyelven (Ács, 2019).

3.5 A szondázó mint fekete doboz modellek elemző eszközzel kapcsolatban sok kritika merült fel (Belinkov, 2021). Több mint 10 ablációs módszert vezettem be, és kimutattam, hogy a morfoszintaktikai szondázó következtetései megbízhatóak.

3.6 A PLM-ek token szintű használata megköveteli az összetett tokenek kezelési módját. Valamilyen aggregáló függvényre van szükség ahhoz, hogy egyetlen vektort kapjunk minden egyes tokenhez. Megmutattam, hogy az aggregáló függvény megválasztása számít, különösen akkor, ha nem tanítjuk tovább a modelleket. 9 aggregáló függvényt hasonlítottam össze 7 nyelven és 3 feladaton (Ács et al., 2021a).

Ezeket az eredményeket Ács (2019); Ács et al. (2021a); Ács et al. (2023) cikkekben publikáltuk. A kísérlet tervezését társszerzőimmel együtt végeztem, az implementáció az én hozzájárulásom.

4. tézis

A morfoszintaktikai szondázó adataim segítségével megmutattam, hogy az egynyelvű PLM-ek jobbak saját nyelvükön, mint a többnyelvű PLM-ek, de a különbség kicsi és gyakran nem statisztikailag szignifikáns. Mind az egynyelvű, mind a többnyelvű PLM-ek sikeresen transzferálhatóak új nyelvekre, amennyiben azok ugyanazt az írásrendszert használják.

Altézisek:

4.1 5 magyar nyelvet támogató PLM-et elemeztem, és kimutattam, hogy a HuBERT (Nemeskey, 2021) – egy kizárólag magyar nyelvű modell – jobb morfológiai, POS- és NER-címkézésben, mint a többnyelvű modellek, különösen a distilBERT, de a különbség kicsi.

4.2 Kiterjesztettem ezt a vizsgálatot az uráli nyelvcsaládra. A magyarhoz hasonló eredményekre jutottam észt és finn modellek eseten (más uráli nyelvekre nincs dedikált modell).

4.3 POS- és NER-címkéző modelleket tanítottam be kisebbségi uráli nyelveken, többnyelvű és egynyelvű modellek transzferálásával. A transzferált modellek rendkívül kis tanulóadat ellenére is sikeresek voltak és state-of-the-art eredményeket értek el nyelvspecifikus erőfeszítések nélkül.

Ezeket az eredményeket a Ács et al. (2021); Ács et al. (2021b) cikkekben publikáltuk. A kísérlet tervezését társszerzőimmel végeztem, az implementáció az én hozzájárulásom.

5. tézis

Tovább finomítottam a szondázási elemzésemet olyan perturbációk segítségével, amelyek célja, hogy meghatározzák a morfoszintaktikai információ pontos helyét egy mondaton belül. Bizonyos információk szisztematikus eltávolítása (perturbációk) feltárja, hogy hol tárolódik az adott információ. A kontextus morfoszintaxisban betöltött szerepének számszerűsítésére Shapley-értékeket használtam, és az eredmények gyakran egybevágnak a nyelvészeti intuíciókkal.

Altézisek:

5.1 Bevezettem perturbációs módszereket, amelyek eltávolítanak bizonyos információkat a mondatból. A morfoszintaktikai szondázó modelleket újratanítottam a 247 szondázó feladaton. Az eredmények betekintést adnak abba, hogy a mondaton belül hol található bizonyos információ. Gyakran találunk ezekre nyelvészeti magyarázatot.

5.2 A perturbációs kísérletek eredményei alapján a nyelvek klaszterezhetők, és az egy nyelvcsaládból származó nyelvek jellemzően egy csoportba kerülnek néhány kivétellel. A nem rokon, de tipológiailag hasonló nyelvek is gyakran egy klaszterbe kerülnek.

5.3 A szondázó mondatokat egy 9 szereplős játékként definiáltam, ahol a játékosok a tokenek a szondázó céltokenhez vett relatív pozíciójuk alapján. Erre alkalmaztam a Shapley keretrendszert.

5.4 Elemeztem a Shapley-értékeket az mBERT, az XLM-RoBERTa és egy LSTM baseline modelleke és azt találtam, hogy a modellek közti hasonlóság magas, ami arra utal, hogy a Shapley-értékek inkább a nyelvről adnak információt, mintsem a modellről. Azonosítottam az átlagtól jelentősen eltérő Shapley-értékeket adó szondázó feladatokat és azt találtam, hogy a Shapley-értékek gyakran esnek egybe a nyelvészeti tulajdonságaival az egyes feladatoknak.

Ezeket az eredményeket (Ács et al., 2023) cikkben publikáltuk. Az implementáció az én munkám.

Kihívások és korlátok

A disszertáció I. része még abból a korszakból származik, amikor a kisebb, konkrét feladatokra specializált neurális modellek voltak jellemzőek, melyeket manapság már ritkábban használnak, különösen új alkalmazások fejlesztésére. Az akkori korlátokat elsősorban az annotált tanítóadatok elérhetősége, illetve egy megfelelő minőségű, szabályalapú morfológiai elemző hiánya jelentette, amely alkalmas lett volna adatok generálására.

A II. rész fő korlátai az általam vizsgált feladatok természetéből adódnak. A morfoszintaktikai feladatok alacsony szintűek, és az egyes modellek teljesítménye ezeken a feladatokon nem feltétlenül írja le, hogy mit várhatunk el tőlük magasabb szintű, sztenderd benchmark feladatokon. Sajnos a nyelvek túlnyomó többsége – még azok is, amelyekhez elegendő nyers szöveg áll rendelkezésre nagyméretű nyelvi modellek tanításához – rendelkezik ilyen benchmarkkal, így azzal kellett dolgoznom, ami elérhető volt.

A II. rész második jelentős korlátja az adatok minősége. Az egyik hibaforrás maga az Universal Dependencies Treebank, amely a fő adatforrásom volt. Bár általánosságban jó minőségű erőforrásnak számít, különféle treebankeket tartalmaz, még ugyanazon nyelven belül is, így elkerülhetetlenek a hibák és következetlenségek. A vizsgált nyelvek magas száma lehetetlenné

tette az egyedi minőségellenőrzést. A második hibaforrás magából a mintavételi eljárásból ered.

A II. rész harmadik fő korlátja a finomhangolás hiánya, néhány kivételtől eltekintve – egy ablációs vizsgálat (3.5. altézis) és az uráli nyelvek kiértékelése (4. tézis). Ennek fő oka az erőforrás-hatékonyság. A finomhangolás a tanítási időt körülbelül nyolcvanszorosára, a kiértékelési időt pedig még ennél is jobban megnövelte volna, mivel minden alkalommal újra be kellett volna tölteni a finomhangolt BERT modelleket.

Gyakorlati alkalmazások

Az 1. tézisben vizsgált kis méretű mélytanulási modellek korlátozott mennyiségű tanítóadat mellett is jó vagy kiváló teljesítményt nyújtanak morfológiai feladatokban. A világ nyelveinek túlnyomó többsége minimális tanítóadattal rendelkezik vagy egyáltalán nem rendelkezik tanítóadattal, és új, annotált adathalmazok létrehozása költséges, vagy gyakran lehetetlen a nyelvész szakértők hiánya miatt. Az olyan enkóder-dekóder modellek, mint amilyeneket vizsgálok, alkalmasak adathalmazok előállításának elindítására (bootstrapping).

A 2. tézis tárgya, a SoPa modell ígéretesnek tűnt mint egy interpretálható neurális modell, amely egyszerű következtetési képességekkel rendelkezik, de jelentősen háttérbe szorult a Transformer-alapú modellek mellett. Így inkább egy érdekes koncepció maradt, kevés gyakorlati alkalmazással.

A disszertációm II. része többnyelvű és egynyelvű modellek széleskörű kiértékelését tartalmazza, ami önmagában is gyakorlati alkalmazásnak számít. Amikor több modell közül kell választani, az egyik legfontosabb tényező a kezdeti értékelés során nyújtott teljesítmény. Az általam kidolgozott morfoszintaktikai szondázási módszertan bármilyen modellre alkalmazható, beleértve a generatív modelleket is, bár különösen jól használható enkóder és enkóder-dekóder architektúráknál. Továbbá, bármely nyelv esetében alkalmazható, amennyiben rendelkezésre áll némi morfológiailag annotált adat.

A 4. tézis két kísérletsorozatot mutat be, amelyek egy-egy nyelvre vagy nyelvcsaládra irányulnak. Ez a fajta összehasonlítás más nyelvekre és szakterületekre is kiterjeszthető. Vizsgálok olyan kevéssé támogatott nyelveket is,

amelyek esetében rendkívül kevés annotált adat áll rendelkezésre, és bemutatom, hogy a modellek jó teljesítménnyel átvihetők (transfer learning). Az így átvitt modellek POS- és NER-címkézőként is használhatók, sok esetben ezek jelentik az egyetlen lehetőséget az alacsony erőforrású nyelvek esetében.

Az 5. tézis a Shapley-értékek finomabb elemzésben való alkalmazhatóságát vizsgálja. Az így kapott elemzési eredmények elegánsak és jól értelmezhetőek, azonban a számítási igény rendkívül nagy, ami jelentős korlátot jelent a széles körű felhasználásban.

Hivatkozások

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *arXiv:2102.12452 [cs]*. ArXiv: 2102.12452.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dávid Márk Nemeskey. 2021. Introducing `huBERT`. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021)*, pages 3–14, Szeged.

Joakim Nivre, Mitchell Abrams, Željko Agić, et al. 2018. Universal Dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Roy Schwartz, Sam Thomson, and Noah A. Smith. 2018. SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines. In *Proc. 56th ACL Annual Meeting*, pages 295–305, Melbourne, Australia.

Kapcsolódó publikációk

Judit Ács. 2018. BME-HAS system for CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 121–126.

Judit Ács. 2019. Exploring BERT's vocabulary. http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html. Accessed: 2021-05-14.

Judit Ács, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and Andras Kornai. 2023. Morphosyntactic probing of multilingual BERT models. *Natural Language Engineering*, page 1–40.

Judit Ács, Ákos Kádár, and Andras Kornai. 2021a. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Judit Ács and András Kornai. 2020. The role of interpretable patterns in deep learning for morphology. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020)*, pages 171–179, Szeged.

Judit Ács, Dániel Lévai, and Andras Kornai. 2021b. Evaluating transferability of BERT models on Uralic languages. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 8–17, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Judit Ács, Dániel Lévai, Dávid Márk Nemeskey, and András Kornai. 2021. Evaluating contextualized language models for Hungarian. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020)*, Szeged.

Ádám Kovács, Judit Ács, András Kornai, and Gábor Recski. 2020. Better together: Modern methods plus traditional thinking in NP alignment. In *Proc. LREC 2020*, pages 3635—-3639.

Ádám Kovács, Evelin Ács, Judit Ács, András Kornai, and Gábor Recski. 2019. BME-UW at SRST-2019: Surface realization with interpreted regular tree grammars. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 35–40, Hong Kong, China. Association for Computational Linguistics.

Attila Nagy, Bence Bial, and Judit Ács. 2021. Automatic punctuation restoration with BERT models. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Gábor Recski, Ádám Kovács, Kinga Gémes, Judit Ács, and Andras Kornai. 2020. BME-TUW at SR'20: Lexical grammar induction for surface realization. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 21–29, Barcelona, Spain (Online). Association for Computational Linguistics.

Gábor Szolnok, Botond Barta, Dorina Lakatos, and Judit Ács. 2021. BME sub-

mission for SIGMORPHON 2021 shared task 0. A three step training approach with data augmentation for morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 268–273, Online. Association for Computational Linguistics.

Egyéb publikációk

Judit Ács. 2014. Pivot-based multilingual dictionary building using Wiktionary. In *The 9th edition of the Language Resources and Evaluation Conference*.

Judit Ács. 2015. Synonym acquisition from translation graph. In *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, pages 14–21. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Judit Ács, Gábor Borbély, Márton Makrai, Dávid Nemeskey, Gábor Recski, and András Kornai. 2018. Hibrid nyelvtechnológiák. *Magyar Tudomány*, 6.

Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. 2015. A two-level classifier for discriminating similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 73–77, Hissar, Bulgaria. Association for Computational Linguistics.

Judit Ács and József Halmi. 2016a. Comparing diacritic restoration methods for Hungarian. In *Proceedings of the Automation and Applied Computer Science Workshop 2016 : AACS'16*. Budapest University of Technology and Economics.

Judit Ács and József Halmi. 2016b. Hunaccent: Small footprint diacritic restoration for social media. In *Proceedings of the First Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*.

Judit Ács and Ágota Illyés. 2015. Language detection and generation. In *Proceedings of the Automation and Applied Computer Science Workshop 2015 : AACS'15*. Budapest University of Technology and Economics.

Judit Ács and András Kornai. 2016. Evaluating embeddings on dictionary-based similarity. In *Proceedings of the First Workshop on Evaluating Vector-Space Representations for NLP (RepEval)*, pages 78–82.

Judit Ács, Dávid Nemeskey, and András Kornai. 2017. Identification of disaster-implicated named entities. In *Proceedings of the First International*

Workshop on Exploitation of Social Media for Emergency Relief and Preparedness.

Judit Ács, Dávid Márk Nemeskey, and Gábor Recski. 2017. Building word embeddings from dictionary definitions. In Katalin Mády Beáta Gyuris and Gábor Recski, editors, *K + K = 120: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS).

Judit Ács, Katalin Pajkossy, and András Kornai. 2013. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria. Association for Computational Linguistics.

Judit Ács and Géza Velkey. 2017. Comparing word segmentation algorithms. In *Proceedings of the Automation and Applied Computer Science Workshop 2017 : AACS'17*. Budapest University of Technology and Economics.

Botond Barta, Endre Hamerlik, Milán Konor Nyist, and Judit Ács. 2025. HuAMR: A Hungarian AMR parser and dataset. In *XXI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2025)*, pages 185–196. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2023. HunSum-1: an abstractive summarization dataset for Hungarian. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, pages 231–243. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2024. From news to summaries: Building a Hungarian corpus for extractive and abstractive summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7503–7509, Torino, Italia. ELRA and ICCL.

Dániel Huszti and Judit Ács. 2017. Entitásorientált véleménykinyerés magyar nyelven. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY*

2017), pages 240–250. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 165–175, Denver, Colorado. Association for Computational Linguistics.

Attila Nagy, Dorina Lakatos, Botond Barta, and Judit Ács. 2023a. TreeSwap: Data augmentation for machine translation via dependency subtree swapping. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 759–768, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Attila Nagy, Dorina Petra Lakatos, Botond Barta, Patrick Nanys, and Judit Ács. 2023b. Data augmentation for machine translation via dependency subtree swapping. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Attila Nagy, Patrick Nanys, Konrád Balázs Frey, Bence Bial, and Judit Ács. 2022. Syntax-based data augmentation for Hungarian-English machine translation. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Gergely Dániel Németh and Judit Ács. 2018. Hyphenation using deep neural networks. In *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Gábor Recski and Judit Ács. 2015. MathLingBudapest: Concept networks for semantic similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 138–142, Denver, Colorado. Association for Computational Linguistics.