Towards the automatic detection of syntactic differences

Kroon, M.S.

Bibliography

Aarts, Flor G. A. M. and Herman Chr. Wekker (1987). *A contrastive grammar of English and Dutch: Contrastieve grammatica Engels / Nederlands*. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-017-4984-8.

Abu-Aisheh, Zeina et al. (2015). 'An exact graph edit distance algorithm for solving pattern recognition problems'. In: *4th International Conference on Pattern Recognition Applications and Methods 2015*.

Abzianidze, Lasha et al. (2017). 'The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 242–247. URL: https://aclanthology.org/E17-2039.

Agić, Željko and Ivan Vulić (2019). 'JW300: A wide-coverage parallel corpus for low-resource languages'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3204–3210. DOI: 10.18653/v1/P19-1310. URL: https://aclanthology.org/P19-1310.

Babická, Blanka et al. (2008). 'The passive voice in English and Czech and some implications for teaching'. In: *Discourse and Interaction* 1.2, pp. 19–30.

Barbiers, Sjef (2009). 'Locus and limits of syntactic microvariation'. In: *Lingua* 119.11, pp. 1607–1623.

Barbiers, Sjef et al. (2005/2008). *Syntactic atlas of the Dutch dialects*. 2 vols. Amsterdam University Press.

Bard, Gregory V. (2007). 'Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric'. In: *Proceedings of the fifth Australasian symposium on ACSW frontiers – Volume 68*. Citeseer, pp. 117–124.

Barron, Andrew, Jorma Rissanen and Bin Yu (1998). 'The minimum description length principle in coding and modeling'. In: *IEEE Transactions on Information Theory* 44.6, pp. 2743–2760.

Benjamini, Yoav and Yosef Hochberg (1995). 'Controlling the false discovery rate: A practical and powerful approach to multiple testing'. In: *Journal of the Royal statistical society: Series B (Methodological)* 57.1, pp. 289–300.

Bertens, Roel, Jilles Vreeken and Arno Siebes (2016). 'Keeping it short and simple: Summarising complex event sequences with multivariate patterns'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 735–744.

Berwick, Robert C. and Noam Chomsky (2016). *Why only us: Language and evolution*. Cambridge, MA: MIT press.

Bonferroni, Carlo (1936). 'Teoria statistica delle classi e calcolo delle probabilità'. In: *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8, pp. 3–62.

Brandsen, Alex (2022). 'Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports'. PhD thesis. Leiden University. URL: `https://hdl.handle.net/1887/3274287`.

Brill, Eric (1992). 'A simple rule-based part of speech tagger'. In: *HLT*.

Broekhuis, Hans (2020). *R-pronominalization and R-words*. Retrieved October 14, 2020 from `https://www.taalportaal.org/taalportaal/topic/link/syntax__Dutch__adp__adp5__P5_strand.xml`.

Brown, Peter F. et al. (1993). 'The mathematics of statistical machine translation'. In: *Computational Linguistics* 19.2, pp. 263–313. URL: `http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf`.

Christodoulopoulos, Christos and Mark Steedman (2015). 'A massively parallel corpus: The Bible in 100 languages'. In: *Language resources and evaluation* 49.2, pp. 375–395.

Cinque, Guglielmo (1999). *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press.

Cinque, Guglielmo and Richard S. Kayne, eds. (2005). *The Oxford handbook of comparative syntax*. Oxford University Press.

Cohen, Jacob (1960). 'A coefficient of agreement for nominal scales'. In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

Cysouw, Michael (2010). 'Semantic maps as metrics on meaning'. In: *Linguistic Discovery* 8.1, pp. 70–95.

Cysouw, Michael and Bernhard Wälchli (2007). 'Parallel texts: Using translational equivalents in linguistic typology'. In: *Language typology and universals* 60.2, pp. 95–99.

Dahl, Östen (2007). 'From questionnaires to parallel corpora in typology'. In: *Language Typology and Universals* 60.2, pp. 172–181.

Davies, Mark (2008). *The sorpus of contemporary American English*. URL: `www.english-corpora.org/coca/`.

de Lange, Joke (2004). 'Article omission in child speech and headlines'. In: *Utrecht Institute of Linguistics OTS Yearbook*, pp. 109–119.

Dempster, Arthur P., Nan M. Laird and Donald B. Rubin (1977). 'Maximum likelihood from incomplete data via the EM algorithm'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Dice, Lee R. (1945). 'Measures of the amount of ecologic association between species'. In: *Ecology* 26.3, pp. 297–302.

Donaldson, Bruce (2008). *Dutch: A comprehensive grammar*. 2nd ed. Comprehensive Grammars. Routledge.

Dušková, Libuše (1991). 'The complex sentence in British and Czech grammar'. In: *Brno studies in English* 19.1, pp. 65–75. URL: http://hdl.handle.net/11222.digilib/104417.

Dyer, Chris, Victor Chahuneau and Noah A. Smith (2013). 'A simple, fast, and effective reparameterization of IBM Model 2'. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648.

Eberhard, David M., Gary F. Simons and Charles D. Fennig, eds. (2021). *Ethnologue: Languages of the World*. 24th ed. Dallas, Texas: SIL International.

Fleiss, J. L. and Jacob Cohen (1973). 'The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability'. In: *Educational and Psychological Measurement* 33, pp. 613–619.

Greenberg, Joseph H. (1963). 'Some universals of grammar with particular reference to the order of meaningful elements'. In: *Universals of Language*. Ed. by Joseph H. Greenberg. Cambridge, MA.: MIT Press, pp. 110–113.

Grünwald, Peter D. (2007). *The minimum description length principle*. MIT press.

Hagberg, Aric A., Daniel A. Schult and Pieter J. Swart (2008). 'Exploring network structure, dynamics, and function using NetworkX'. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Ed. by Gäel Varoquaux, Travis Vaught and Jarrod Millman. Pasadena, CA USA, pp. 11–15.

Haspelmath, Martin (1997). *Indefinite pronouns*. Oxford University Press.

— (2003). 'The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison'. In: *The new psychology of language*. Psychology Press, pp. 217–248.

Hinrichs, Frauke and Jilles Vreeken (2017). 'Characterising the difference and the norm between sequence databases'. In: *Proceedings of the 4th Workshop on Interactions between Data Mining and Natural Language Processing (DMNLP)*.

Holmberg, Anders (2016). 'Null subjects in Finnish and the typology of pro-drop'. In: *Uralic Syntax book project. Cambridge: Cambridge University*.

Jalili Sabet, Masoud et al. (2020). 'SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings'. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 1627–1643. DOI: 10.18653/v1/2020.findings-emnlp.147. URL: https://aclanthology.org/2020.findings-emnlp.147.

Kiss, Katalin É. (2002). *The syntax of Hungarian*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511755088.

Koehn, Philipp (2005). 'Europarl: A parallel corpus for statistical machine translation'. In: *MT summit*. Vol. 5, pp. 79–86.

Kroon, Martin et al. (2019). 'A filter for syntactically incomparable parallel sentences'. In: *Linguistics in the Netherlands* 36. Ed. by Janine Berns and Elena Tribushinina, pp. 147–161. DOI: https://doi.org/10.1075/avt.00029.kro.

— (2020). 'Detecting syntactic differences automatically using the Minimum Description Length principle'. In: *Computational Linguistics in the Netherlands Journal* 10, pp. 109–127.

Ladányi, Mária (2015). 'Particle verbs in Hungarian'. In: *Word-formation: An international handbook of the languages of Europe*. Ed. by Peter O. Müller et al. Vol. 1. De Gruyter Mouton, pp. 660–672.

Levenshtein, Vladimir I. (1966). 'Binary codes capable of correcting deletions, insertions, and reversals'. In: *Soviet physics doklady* 10.8, pp. 707–710.

Lin, Chih-Long (1994). 'Hardness of approximating graph transformation problem'. In: *Algorithms and Computation*. Ed. by Ding-Zhu Du and Xiang-Sun Zhang. Vol. 843. Lecture Notes in Computer Science. Berlin: Springer, pp. 74–82.

Lison, Pierre and Jörg Tiedemann (2016). 'Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles'. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Malá, Markéta (2014). *English copular verbs: A contrastive corpus-supported view*. Filozofická fakulta Univerzity Karlovy.

Mårdh, Ingrid (1980). *Headlinese: On the grammar of English front page headlines*. Vol. 58. Liberläromedel/Gleerup.

McNemar, Quinn (1947). 'Note on the sampling error of the difference between correlated proportions or percentages'. In: *Psychometrika* 12.2, pp. 153–157.

Naughton, James (2005). *Czech: An essential grammar*. Essential grammars. Routledge.

Nerbonne, John and Wybo Wiersma (2006). 'A measure of aggregate syntactic distance'. In: *Proceedings of the Workshop on Linguistic Distances*. Association for Computational Linguistics, pp. 82–90.

Nivre, Joakim et al. (2016). 'Universal Dependencies v1: A Multilingual Treebank Collection'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), pp. 1659–1666. URL: https://aclanthology.org/L16-1262.

Och, Franz Josef and Hermann Ney (2003). 'A Systematic Comparison of Various Statistical Alignment Models'. In: *Computational Linguistics* 29.1, pp. 19–51.

Odijk, Jan et al. (2017). 'The parse and query (PaQu) application'. In: *CLARIN in the Low Countries*. Ed. by Jan Odijk and A. Hessen. London: Ubiquity Press, pp. 281–297. DOI: https://doi.org/10.5334/bbi.23.

Oostdijk, Nelleke et al. (2013). 'The construction of a 500-million-word reference corpus of contemporary written Dutch'. In: *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*. Springer Verlag. Chap. 13.

Osborne, Miles (1999a). 'DCG induction using MDL and parsed corpora'. In: *International Conference on Learning Language in Logic*. Springer, pp. 184–198.

— (1999b). 'MDL-based DCG induction for NP identification'. In: *EACL 1999: CoNLL-99 Computational Natural Language Learning*.

Östling, Robert and Jörg Tiedemann (2016). 'Efficient word alignment with Markov Chain Monte Carlo'. In: *Prague Bulletin of Mathematical Linguistics* 106, pp. 125–146. URL: http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf.

Radford, A. (2004). *English syntax: An introduction*. Cambridge University Press. ISBN: 9780521542753. URL: https://books.google.nl/books?id=LdAi292Q4-0C.

Reback, Jeff et al. (Mar. 2021). *pandas-dev/pandas: Pandas 1.2.3*. Version v1.2.3. DOI: 10.5281/zenodo.4572994. URL: https://doi.org/10.5281/zenodo.4572994.

Rounds, Carol (2009). *Hungarian: An essential grammar*. Routledge.

Sampson, Geoffrey (2000). 'A proposal for improving the measurement of parse accuracy'. In: *International Journal of Corpus Linguistics* 5.1, pp. 53–68.

Sanders, Nathan C. (2007). 'Measuring syntactic difference in British English'. In: *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*. Association for Computational Linguistics, pp. 1–6.

— (2010). 'A statistical method for syntactic dialectometry'. PhD thesis. Indiana University.

Sanguinetti, Manuela and Cristina Bosco (2015). 'PartTUT: The Turin University Parallel Treebank'. In: *Harmonization and development of resources and tools for Italian natural language processing within the PARLI project*. Springer, pp. 51–69.

Shannon, Claude E. (1948). 'A mathematical theory of communication'. In: *The Bell system technical journal* 27.3, pp. 379–423.

Spruit, Marco René (2008). 'Quantitative perspectives on syntactic variation in Dutch dialects'. PhD thesis. University of Amsterdam.

Straka, Milan and Jana Straková (2017). 'Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe'. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, pp. 88–99. URL: http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Tatti, Nikolaj and Jilles Vreeken (2012). 'The long and the short of it: Summarising event sequences with serial episodes'. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 462–470.

Taylor, Ann, Mitchell Marcus and Beatrice Santorini (2003). 'The Penn treebank: An overview'. In: *Treebanks*, pp. 5–22.

Tiedemann, Jörg (2011). 'Bitext alignment'. In: *Synthesis Lectures on Human Language Technologies* 4.2, pp. 1–165.

— (2012). 'Parallel data, tools and interfaces in OPUS'. In: *LREC*. Vol. 2012. Citeseer, pp. 2214–2218.

Toutanova, Kristina et al. (2003). 'Feature-rich part-of-speech tagging with a cyclic dependency network'. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 252–259.

Van Craenenbroeck, Jeroen, Marjo van Koppen and Antal van den Bosch (2019). 'A quantitative-theoretical analysis of syntactic microvariation: Word order in Dutch verb clusters'. In: *Language* 95.2, pp. 333–370.

van den Bosch, Antal et al. (2007). 'An efficient memory-based morphosyntactic tagger and parser for Dutch'. In: *LOT Occasional Series* 7, pp. 191–206.

van der Klis, Martijn, Bert Le Bruyn and Henriette De Swart (2017). 'Mapping the PERFECT via translation mining'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Vol. 2, pp. 497–502.

Wälchli, Bernhard (2007). 'Advantages and disadvantages of using parallel texts in typological investigations'. In: *Language Typology and Universals* 60.2, pp. 118–134.

— (2010). 'Similarity semantics and building probabilistic semantic maps from parallel texts'. In: *Linguistic Discovery* 8.1, pp. 331–371.

Weir, Andrew (2009). 'Article drop in English headlinese'. MA thesis. University College London.

Wiersma, Wybo, John Nerbonne and Timo Lauttamus (2011). 'Automatically extracting typical syntactic differences from corpora'. In: *Literary and Linguistic Computing* 26.1, pp. 107–124.

Wong, Tak-sum et al. (2017). 'Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank'. In: *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Pisa,Italy: Linköping University Electronic Press, pp. 266–275. URL: https://www.aclweb.org/anthology/W17-6530.

Youden, William J. (1950). 'Index for rating diagnostic tests'. In: *Cancer* 3.1, pp. 32–35.

Zeng, Zhiping et al. (2009). 'Comparing stars: On approximating graph edit distance'. In: *Proceedings of the VLDB Endowment* 2.1, pp. 25–36.

Zwart, Jan-Wouter (2011). *The syntax of Dutch*. Cambridge Syntax Guides. Cambridge University Press. DOI: 10.1017/CBO9780511977763.

Overview of URLs to used, referenced and developed tools and datasets

Detecting Syntactic Differences Automatically — *This dissertation*

General link:
`https://github.com/mskroon/DeSDA`

Bible corpus — (Christodoulopoulos and Steedman 2015)

General link:
`https://github.com/christos-c/bible-corpus`

DITTO — (Bertens, Vreeken and Siebes 2016)

General link:
`http://eda.mmci.uni-saarland.de/prj/ditto`

`eflomal` — (Östling and Tiedemann 2016)

General link:
`https://github.com/robertostling/eflomal`

Europarl v7 corpus — (Koehn 2005)

General link:
`https://www.statmt.org/europarl`

`fast_align` (Dyer, Chahuneau and Smith 2013)

General link:

`https://github.com/clab/fast_align`

Frog tagger (van den Bosch et al. 2007)

General link:

`http://languagemachines.github.io/frog`

GIZA++ (Och and Ney 2003)

General link:

`https://www.statmt.org/moses/giza/GIZA++.html`

`networkx` (Hagberg, Schult and Swart 2008)

General link:

`https://networkx.org`

Download:

`https://pypi.org/project/networkx`

Opus corpus (including Europarl v7) (Tiedemann 2012)

General link:

`https://opus.nlpl.eu`

`pandas` (Reback et al. 2021)

General link:

`https://pandas.pydata.org`

Download:

`https://pypi.org/project/pandas`

SimAlign (Jalili Sabet et al. 2020)

General link:

`https://github.com/cisnlp/simalign`

SQS (Tatti and Vreeken 2012)

General link:

`http://adrem.uantwerpen.be/sqs`

Stanford tagger (Toutanova et al. 2003)

General link:

`https://nlp.stanford.edu/software/tagger.shtml`

UDPipe (Straka and Straková 2017)

General link:

`https://ufal.mff.cuni.cz/udpipe`

Tool:

`https://github.com/ufal/udpipe`

Models (from 15 Nov 2018; used in Chapters 2 and 3):

`https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2898`

Model (English ParTUT; used in Chapter 4):

`https://github.com/UniversalDependencies/UD_English-ParTUT`

Universal Dependencies (Nivre et al. 2016)

General link:

`https://universaldependencies.org`