Faculty of Informatics
Masaryk University

Statistical Processing of Text Corpora

Pavel Rychlý

Habilitation Thesis
(Collection of Articles)

March 2015

Abstract

Availability of huge corpora changed significantly several research fields. A relatively new "computational linguistics" provides quite different results from "original" linguistics, different both in form and quality. Computational linguistics can create resources helpful for processing natural languages on computers. It can also provide much bigger and detailed analysis of language phenomena. In the field of lexicography, today, new high quality dictionaries are always created with the help of a computer analysis of a big corpus or corpora.

Behind all these changes, there is some kind of processing of large texts and most of it is statistical processing. This habilitation thesis contains a short introduction to statistical processing of texts and lists ten papers in this field. The papers were selected in three areas: collocations, distributional semantics and language modeling.

The most important paper of this collection was presented at the Euralex conference (the biggest lexicography conference) in 2004: *The Sketch Engine*. It describes a new software for linguists and especially lexicographers. After more than ten years, the system is well known world-wide and considered as a standard tool in lexicography. The paper has hundreds of citations, the system has thousands of active users.

The system also forms a hub for other papers in the collection, most of them describe a component incorporated into the system or some part of the system was used during a research finished by the paper. The author of this thesis is the original designer and programmer of the whole system and he is the head of the team behind the current development of the system.

Abstrakt

Dostupnost velkých korpusů podstatným způsobem změnila několik vědních oborů. Relativně nová ,,počítačová lingvistika" poskytuje výrazně jiné výsledky než ,,původní" lingvistika, jiné jak ve formě tak v kvalitě. Počítačová lingvistika dokáže vytvářet zdroje, které jsou dobře využitelné při počítačovém zpracování přirozených jazyků. Poskytuje též mnohem větší a detailnější analýzu ruzných jazykových jevů. V oblasti lexikografie dnes nové kvalitní slovníky vznikají vždy s podporou počítačové analýzy velkého korpusu či několika korpusů.

Za všemi těmito změnami stojí nějaký druh zpracování textů a ve většině případů jde o statistické zpracování. Tato habilitační práce obsahuje úvod do statistického zpracování textů a dále deset článků k tomuto tématu. Články byly vybrány z následujících oblastí: kolokace, distributivní sémantika a jazykové modelování.

Nejdůležitější článek z tohoto souboru byl prezentována na konferenci Euralex (nejvýznamnější lexikografická konference) v roce 2004, nese název *,,The Sketch Engine"*. Popisuje nový systém pro lingvisty a zejména lexikografy. Po více než deseti letech je systém známý po celém světě a je považován za standard mezi nástroji pro lexikografy. Článek má stovky citací, systém má tisíce aktivních uživatelů.

Systém je také společným jmenovatelem pro ostatní články z předloženého souboru, většina z nich popisuje nějakou komponentu začleněnou do uvedeného systému nebo nějaká část systému byla použita během výzkumu završeném zde uvedeným článkem. Autor této habilitační práce původně navrhl a naprogramoval celý systém a v současné době je vedoucím týmu, který systém dále vyvíjí.

Contents

Chapter 1

Statistical Processing of Text Corpora: Commentary

1.1 Introduction

Natural language processing was one of the computer applications from the very beginning of electronic computers. Using statistics for natural language processing is one of two generic methods. The second one is rule based approach. Statistical approach requires some data to train on, on the other hand, rule based approach needs knowledge of an expert (linguist) that is transformed into some formal rules. In many applications rule based and statistical methods are combined together because some parts of natural language are well understood or even formally described by linguists but other parts are described only by examples.

1.1.1 Text Corpora

Text corpora are big collections of texts in a uniform format usually with some additional annotation. Such corpora are natural sources of data for any statistical processing of a language. The first corpus in the modern form (Brown Corpus [FK64]) was developed in 60s of the last century and it contains one million of words in 500 texts. It was a big step at that times, but it is quite small as representation of English. It can be (and was) used for computing global characteristics of English: average length of words/sentences, relative frequency of common words (articles, prepositions, etc.), frequency of part of speech and so on [KF67]. Any corpus of such size, however, cannot be used for gaining any information about individual words, because there are only several occurrences of them. On the other hand, the annotated version is still used for training and evaluating part-of-speech taggers [SV97].

The first big corpus (British National Corpus [Bur95] and in the same time also the Bank of English [Jär94]) was prepared in early 90s, it is more than 100 times bigger and it was developed by a consortium of dictionary publishers and universities. In such corpus we can find information about thousands of most frequent words and lexicographers used it for building better modern dictionaries. Today, we have another 100 times bigger cor-

pora (containing 10 billion words) in which even rare words or phrases have enough occurrences to explore. That is much more data than we ever had.

For an illustration of the size of such corpora, we can consider how much text people can read. With average reading speed 200 words per minute, one can read 80 million words during one year of 18 hours reading each day. One billion words mean reading 4 hours every day for about 50 years. It is clear that most of native speakers of any language have not read more than 1 billion words yet. And probably no one or only a few people have read more than 10 billion of words in their whole life.

The turn of the century bring several good text books covering statistical processing of texts, most notably [MS99, JM00]. That was also time of first big corpora created from the Web [Sha04, FZBB08], later leading to huge corpora with more than 70 billion words [PJR12]. Several researchers [Chu11] think that today preference of statistical approaches is too strong and we should try to use more linguistics/rules in future research or systems.

Both statistical and rule based worlds have made a big progress in the last several decades with the use of big text corpora. Statistical methods can exploit much bigger data for training, we can train much more features from bigger data. For linguists, corpora provide a big inventory of real language uses and they can be used to test linguistic theories and form better understanding how a natural language works. For example, popularity on the open market proves that corpus based dictionaries have higher quality.

1.1.2 Presented Topics

This thesis is a collection of ten papers. They were selected to match the overall topic of this thesis and the author of this thesis is the only or the main author of all but one of them.

The papers are grouped into three related topics. The first five papers deal with collocations, usually one of the first aggregated characteristics of a word in a corpus. Next two papers are about building and using distributional thesaurus which is based on similarity of collocations for different words or lemmas. Last three papers explore basics of language modeling that try to find most probable word sequences from a list of candidates.

The following sections describe each of the three selected topics in more details.

1.2 Collocations

Collocations of a word are words (or lemmas or any other corpus element) which occur in the context of a given word frequently. Such definition brings questions: what is a context, what does it mean 'frequently'. Context is usually the previous or following word, or a window of words (for example: from 5 words left to 3 words right). Later we will see that more fine-grained context definition could provide much better results.

As 'frequency', we can really use number of occurrences. In such case, most frequent collocations for almost any word in English are "the" and "a", that is not very interesting. That is the reason why more sophisticated

statistical measures are used in many tools [SC94, Sma94, Dav09, Ant04, Har12].

1.2.1 Collocation statistical measures

There are many statistical association measures used to identify good collocations, T-score [CGHK91] and MI-score [CH90] are widely used, quite exhaustive list of used measures is available in [Eve08]. Most of these measures define a formula of an association score which indicates amount of statistical association between two words.

There are two general problems of most association scores:

1. A score is fine-tuned to one particular corpus size and/or key word frequency. If we use a score for a corpus with very different number of tokens the resulting list is not satisfying enough or is completely wrong.

2. The score values are usually meaningless and corpus specific, they cannot be used for comparing words (or word pairs) of different corpora. But end-users want an interpretation of such scores and want a score's stability. They want to compare collocation scores of different words and on different corpora or subcorpora.

A new measure logDice was defined in [Ryc08] (see Section 1.5.1 and Chapter 2), based on the Dice coefficient [Dic45], it has a reasonable interpretation, scales well to varying corpus size, is stable on subcorpora, and the values are in a reasonable range.

1.2.2 Filtering

Another way to deal with the *"the-is-the-best-collocation"* problem is to incorporate some rules with linguistic knowledge. Using a stop list containing most frequent words (articles, prepositions etc.) or filtering specific part of speech (adjectives/nouns) works very well.

More advanced linguistic knowledge include using a parser to specify the context [SW06]. In such case we define, for example, *adjective modifier* as a context of *noun*. That is also the key feature of word sketches in the Sketch Engine [KRST04] (see Section 1.5.3 and Chapter 4), where full syntactic parsing is substituted by shallow parsing using corpus queries.

1.3 Distributional Semantics

Semantics is one of the higher levels of natural language processing. Handling semantics or meaning of sentences or documents is very complicated and there is no clear consensus what the result should look like. On the other hand, finding a meaning of one word looks like a clear task: there is a list of meanings of the given word in some inventory (dictionary) and we should find the best meaning for the given word in the given context. It is important to know that we need the given context because without

it words are usually meaningless. Unfortunately, the task is clear but the results are very disappointing. Human annotators are not able to find correct meanings in many contexts or the inter-annotator agreement is very low [YF99, HMP+06, SOJN08]. The problem is that while dictionaries list meanings as disjoint items, in reality a word's meaning is defined by its context: words have meaning potentials, a word could have several meanings (or a combinations of meanings) in one context [Han00, Han13]. Distributional Semantics is an attempt to overcome that problem. We do not work with discrete meanings, we work with some probabilities of similarity.

With distributional semantics we are trying to find most similar words (or phrases) for a given word together with respective probabilities (or scores). The biggest advantage of this approach is much wider coverage of the lexicon; we can compute similarity for any word (lemma) or even a multi-word (phrase). Such data could be used in many different applications, for example [Ryc14] (see Section 1.5.7 and Chapter 8, more examples in [PCB+09]). Depending on chosen context variant (again preceding/following word, words window, grammatical relations, etc.) and similarity measure [Lee99, WWM04] the resulting data are suitable for different purposes [VdPT06, Wee03].

1.3.1 Computing Distributional Semantics

The result of computing distributional semantics from a big corpus is some form of thesaurus for all words with number of occurrences above some threshold. This threshold could also be defined on number of different context the word occurs in, because there could be words (for example parts of names) with high frequency but without any match in a grammatical relation [RK07].

The main problem of using any form of such thesaurus is its size. There could be gigabytes of data which have to be processed with some form of random access. Much bigger problem is how to compute the whole thesaurus. Early works reported long computation times [GC06], but there are approaches (usually based on some form of MapReduce [DG08]) which are fast enough even for many-billion word corpus [Ryc14] (see Section 1.5.6 and Chapter 7)

1.4 Language Modeling

Language modeling is part of natural language processing which can use high levels of language analysis (up to semantics) but it could be used on lower levels (down to character or speech recognition). There are two main tasks for a language model:

1. compute the probability of an input,
2. find most probable input for a given output (together with noisy channel model).

Most language models are based on some form of conditional probabilities, computed from large text corpora. During last years there is a rise of language models based on neural networks [Mik12, CW08].

We have done several experiments with probabilistic language models [Ryc11] and one application which uses a language model [Ryc12].

1.5 Contributions

Author of this thesis together with several coauthors published more than 40 articles. This thesis is a collection of ten of them. They were selected to match the overall theme of this thesis and the author of this thesis is the only or the main author of all but one of them.

Each of the following chapters contains one article. Four of the them describe a software system as a whole, the rest of the papers present some key features of these systems. Each of the systems is used by hundreds or thousands of users. One article was published in a journal, others were originally presented on a conference.

1.5.1 A lexicographer-friendly association score

This paper reacts to the problem of meaningless association score values of most association measures used in corpus linguistics. It defines a new score *logDice* which has a reasonable interpretation and its values do not depend on the corpus size, hence the scores could be directly compared on corpora and subcorpora with different sizes.

The logDice score is build in as the default score for finding collocations and sorting items of word sketches in the Sketch Engine.

1.5.2 Manatee/Bonito – A Modular Corpus Manager

This paper describes main features of the Manatee corpus management system including graphical user interfaces Bonito and Bonito2. Today, both Manatee and Bonito are distributed under open source license (GNU GPL) as NoSketch Engine package from Natural Language Processing Centre web site (`http://nlp.fi.muni.cz/trac/noske/`) The system is in regular use by many research groups at universities around the world. There are also commercial companies (especially publishers) which use the system or its parts or extensions in day-to-day works.

1.5.3 The Sketch Engine

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. They were presented before, but they only existed for English. This paper describes a new corpus tool which generates word sketches and is language independent and tagset independent. For each language it requires definitions of grammatical relations in the form of corpus queries on part of speech tags.

The Sketch Engine is both product and service. As a product, it could be installed on a computer and used for a local corpus indexing and exploration. There are several tens of such local installations in universities and companies (mostly publishing houses) around the world. As a service, it is accessible on the Sketch Engine web server with hundreds of pre-installed corpora in more than 70 languages. There are thousands of regular users both academic and commercial performing hundreds of thousands of operations [Buš14].

The paper was published in 2004 at EURALEX – the biggest world conference on lexicography. Later it was published as a technical report ITRI-04-08 at Information Technology Research Institute, University of Brighton. The paper was also selected for publication as a chapter in the book *Practical Lexicography: A Reader* published by Oxford University Press in 2008.

The author of this thesis have done the whole design and implementation of the system. He also worked on the evaluation described in the paper.

1.5.4 The Sketch Engine: ten years on

The Sketch Engine was born in 2004 by Lexical Computing Ltd. The system was (and still is) further developed. This paper describes the core functions (word sketches, concordancing, thesaurus) and also many new features during the ten years of development: Good Dictionary Examples [KHM+08, KHM11], learner corpora [KBK+13], bilingual sketches [BJK+14], terminology finding [KJK+14].

The author of this thesis is the original author of the system and the leader of the development team.

1.5.5 Behaviour of Collocations in the Language of Legal Subdomains

This paper examines the collocational behaviour of multi-word expressions in legal sublanguages. It is an application of collocational analysis on a domain text. It compares language of primary regulations (statutory law) with language of secondary regulations (government decrees). The paper shows that those sublanguages are quite different.

The author of this thesis has done most of the data preparation and analysis, and also writing the text.

1.5.6 An efficient algorithm for building a distributional thesaurus

Creating a distributional thesaurus requires huge amount of computation. A direct approach looks at each word and compares it with each other word, checking all contexts to see if they are common. Thus, complexity is $O(n^2m)$ where n in the number of types and m is the size of the context vector. Both n and m could be hundreds of thousands or even millions for a billion word corpus. These numbers led to published estimates that full calculation will take nearly 300 days [GC06]. Our paper proposes a method how to compute such thesaurus in under 2 hours. The algorithm is incorporated into the

Sketch Engine and the paper also presents another innovative development in the same system.

The author of this thesis is the designer of the algorithm and the programmer of the system. He also work on the evaluation and the text writing.

1.5.7 Finding the Best Name for a Set of Words Automatically

Many natural language processing applications use clustering or other statistical methods to create sets of words. Such sets group together words with similar meaning and in many cases humans can find an appropriate term quickly. On the other hand computers represent such sets with a meaningless number or ID.

The paper proposed an algorithm for finding names for a set of words. The proposed method exploits the distributional thesaurus data which provide a list of similar words for a given word. The implementation is mostly language and corpus independent and works quite well for many test data.

1.5.8 CzAccent – Simple Tool for Restoring Accents in Czech Texts

There are many Czech texts written without any accents. The paper describes a tool for fully automatic restoration of Czech accents. The system is based on a simple approach of a big lexicon. The resulting accuracy of the system evaluated on large Czech corpora is quite high.

The algorithm behind is based on unigram word model, but because of precise construction of the language model data it was one of the best systems for adding accents in Czech texts. Even after more than 10 years from the CzAccent creation, it is still one of the fastest systems.

The tool is accessible on the Natural Language Processing Centre website `http://nlp.fi.muni.cz/cz_accent/`, it is in regular use by hundreds of users from around the whole world.

1.5.9 Frequency of Low-Frequency Words in Text Corpora

Low-frequency words, esp. words occurring only once in a text corpus, are very popular in text analysis. Also many lexicographers draw attention to such words.

This paper is one of the author's experiments with fundamentals of language modelling: how to estimate (almost) non-visible item probabilities from text corpora. The paper lists a detailed statistical analysis of low-frequency words. The results provides important information for many practical applications, including lexicography and language modeling.

1.5.10 Words' Burstiness in Language Models

Good estimation of the probability of a single word is a crucial part of language modelling. It is based on raw frequency of the word in a training corpus. Such computation is a good estimation for functional words and most very frequent words, but it is a poor estimation for most content words because of words' tendency to occur in clusters. This paper provides an analysis of words' burstiness and propose a new unigram language model which handles bursty words much better. The evaluation of the model on two data sets shows consistently lower cross-entropy in the new model.

Bibliography

[Ant04] L Anthony. Antconc: A learner and classroom friendly. *Multi-Platform Corpus*, 2004.

[BJK+14] Vít Baisa, Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, and Pavel Rychlý. Bilingual word sketches: the translate button. In *Proc EURALEX*, 2014.

[Bur95] Lou Burnard. The BNC reference manual, 1995.

[Buš14] Jan Bušta. SkE infrastructure. Presented at XVI EURALEX International Congress, Bolzano, Italy, 2014.

[CGHK91] K. Church, W. Gale, P. Hanks, and D. Kindle. 6. Using Statistics in Lexical Analysis. *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 1991.

[CH90] K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[Chu11] Kenneth Church. A pendulum swung too far. *Linguistic Issues in Language Technology*, 6, 2011.

[CW08] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[Dav09] Mark Davies. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190, 2009.

[DG08] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[Dic45] L.R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 1945.

[Eve08] Stefan Evert. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2, 2008.

[FK64] W Nelson Francis and Henry Kucera. Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island*, 1, 1964.

[FZBB08] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

[GC06] James Gorman and James R Curran. Scaling distributional similarity to large corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 361–368. Association for Computational Linguistics, 2006.

[Han00] Patrick Hanks. Do word meanings exist? *Computers and the Humanities*, 34(1):205–215, 2000.

[Han13] Patrick Hanks. *Lexical analysis: Norms and exploitations*. Mit Press, 2013.

[Har12] Andrew Hardie. Cqpweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, 17(3):380–409, 2012.

[HMP+06] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics, 2006.

[Jär94] Timo Järvinen. Annotating 200 million words: the bank of english project. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 565–568. Association for Computational Linguistics, 1994.

[JM00] Dan Jurafsky and James H Martin. *Speech & language processing*. Pearson Education India, 2000.

[KBK+13] Iztok Kosem, Vít Baisa, Vojtěch Kovář, Adam Kilgarriff, et al. User-friendly interface of error/correction-annotated corpus for both teachers and researchers. In *Learner Corpus Research.*, 2013.

[KF67] Henry Kucera and Nelson Francis. *Computational analysis of present-day American English*. Brown university press, 1967.

[KHM+08] A. Kilgarriff, M. Husák, K. McAdam, M. Rundell, and P. Rychlý. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proceedings of the XIIIth EURALEX International Congress. Barcelona: Universitat Pompeu Fabra*, pages 425–432, 2008.

[KHM11] Iztok Kosem, Milos Husak, and Diana McCarthy. Gdex for slovene. In *Electronic lexicography in the 21st century: New Applications for New Users: Proceedings of eLex 2011, Bled, 10-12 November 2011*, pages 151–159, 2011.

[KJK+14] Adam Kilgarriff, Miloš Jakubíček, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. Finding terms in corpora for many languages with the sketch engine. In *Proceedings of the Demonstrations at the 14th Conferencethe European Chapter of the Association for Computational Linguistics*, pages 53–56, Gothenburg, Sweden, 2014. The Association for Computational Linguistics.

[KRST04] A. Kilgarriff, P. Rychlý, P. Smrž, and D. Tugwell. The Sketch Engine. *Proceedings of Euralex*, pages 105–116, 2004.

[Lee99] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 1999.

[Mik12] Tomáš Mikolov. *Statistical Language Models Based on Neural Networks*. PhD thesis, Ph. D. thesis, Brno University of Technology, 2012.

[MS99] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[PCB+09] Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. Web-scale distributional similarity and entity set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 938–947. Association for Computational Linguistics, 2009.

[PJR12] Jan Pomikálek, Milos Jakubícek, and Pavel Rychlý. Building a 70 billion word corpus of english from clueweb. In *LREC*, pages 502–506, 2012.

[RK07] Pavel Rychlý and Adam Kilgarriff. An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 41–44. Association for Computational Linguistics, 2007.

[Ryc08] Pavel Rychlý. A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9, 2008.

[Ryc11] Pavel Rychlý. Words' burstiness in language models. *RASLAN 2011 Recent Advances in Slavonic Natural Language Processing*, page 131, 2011.

[Ryc12] Pavel Rychlý. Czaccent–simple tool for restoring accents in czech texts. *RASLAN 2012 Recent Advances in Slavonic Natural Language Processing*, page 85, 2012.

[Ryc14] Pavel Rychlý. Finding the best name for a set of words automatically. *RASLAN 2014 Recent Advances in Slavonic Natural Language Processing*, page 77, 2014.

[SC94] B. Schulze and O. Christ. The IMS Corpus Workbench. *Institut fur maschinelle Sprachverarbeitung, Universitat Stuttgart*, 1994.

[Sha04] Serge Sharoff. Analysing similarities and differences between corpora. In *Proceedings of the 7th Conference of Language Technologies (Jezikovne Tehnologije)*, volume 83, 2004.

[Sma94] F.Z. Smadja. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1):143–177, 1994.

[SOJN08] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[SV97] Christer Samuelsson and Atro Voutilainen. Comparing a linguistic and a stochastic tagger. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 246–253. Association for Computational Linguistics, 1997.

[SW06] Violeta Seretan and Eric Wehrli. Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 953–960. Association for Computational Linguistics, 2006.

[VdPT06] Lonneke Van der Plas and Jörg Tiedemann. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 866–873. Association for Computational Linguistics, 2006.

[Wee03] Julie Elizabeth Weeds. *Measures and applications of lexical distributional similarity.* PhD thesis, University of Sussex, 2003.

[WWM04] Julie Weeds, David Weir, and Diana McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 10–15. Association for Computational Linguistics, 2004.

[YF99] Chung Yong and Shou King Foo. A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99)*, 1999.