CURRICULUM VITAE
Gideon (Jozua) Kotzé

7 21st Street, Menlo Park, Pretoria, 0081
Email: gidi8ster@gmail.com
Website: www.gideonkotze.nl

PERSONAL INFORMATION
Nationality: South African
Date of birth: 9 June 1981
Languages: In order of proficiency: Afrikaans (mother tongue), English, Dutch, German and French, limited knowledge of Xhosa (high school education) and Japanese (spoken language course)

WORKING EXPERIENCE

2014-   Postdoctoral Fellow at University of South Africa (UNISA)

2012   Designed and implemented a system using transformation-based learning to correct and apply high accuracy tree-to-tree alignments between phrase-structure trees. The goal is the creation of large-scale parallel treebanks for syntax-based machine translation, as with the PaCo-MT project (see below). A paper on this work was published in the Computational Linguistics in the Netherlands (CLIN) Journal of 2012. This work was done as part of the research done for my Ph.D project (see Education).

2009   Taught a Bachelors level course at the University of Groningen "Tekstmanipulatie" (Text Manipulation) with Dr Barbara Plank.

2008-2011   I worked on the machine translation project Parse and Corpus-Based Machine Translation (PaCo-MT) with the Catholic University of Leuven, Belgium, and the company OneLiner bvba in Sint Niklaas, Belgium. The other members of the consortium consisted of Frank van Eynde, Vincent Vandeghinste, Jörg Tiedemann, Joachim van den Bogaert and Koen Desmet. My main task consisted of the finding or development and subsequent application of tools for the production of large parallel treebanks for use in the MT system. The main technologies applied in this task were tokenisation, sentence alignment, word alignment, syntactic parsing and tree-to-tree alignment. The concerned language pairs were Dutch to French, French to Dutch, Dutch to English, and English to Dutch.

2008   Contract work completed for the Centre for Text Technology (CTexT) in Potchefstroom, South Africa. Main activity comprised the extension of the abovementioned Afrikaans wordnet to a total of 10 068 entries, in which English entries were mostly translated into Afrikaans and some data manipulation was done by programming in Perl.

2006-2007   Temporary position as data developer and researcher at the Centre for Text Technology (CTexT) at the Potchefstroom campus of the North-West University, with Professor Gerhard van Huyssteen as the head of the centre. Activities included: morphological analysis and quality control, data manipulation by programming in Perl, the design of databases (specifically *ALEXANDER* (Afrikaans Lexicon and Annotated Database for Engineering and Research)), linguistic examination and quality control of lexica, corpus analysis and translation.

2006	Internship at the Vrije Universiteit in Amsterdam as part of M.A. training. Position: Fellow worker of a project called *Cornetto* (Combinatorial and Relational Network as Toolkit for Dutch Language Technology) , an electronic lexical database in Dutch for human language technological purposes. Various institutions are involved. The activities included semantic annotation and data manipulation by programming.
2002-2004	Computer assistant/consultant at the former University of Port Elizabeth in computer labs for general use.

COMPUTER SKILLS

Text production and relevant skills

Microsoft Office (Excel, Word, Access, Powerpoint), OpenOffice (the text, spreadsheet and presentation features), LaTeX. I also have experience using various editors such as Emacs, Nano, Kate, Kwrite, TexShop, TexMaker and Sublime Text 2, for both text production and programming.

Programming

Perl (self-taught, main language of choice for the PaCo-MT project and my PhD research), Bash (also used in this project and research), C++ and Java (Bachelor level courses), a basic knowledge of R (self-taught, applied for PhD research), Turbo Pascal (secondary school course). I acquired some knowledge of XHTML, JavaScript, VBScript and ASP as part of a Web Technology course on Bachelor level. Some experience with databases (Microsoft Access, SQL). As a means of representing data I have much experience in the use of XML. Finally, I have experience using the version control software Git.

Computational linguistics tools

Moses and GIZA++ (statistical machine translation and word alignment tools), syntactic parsers (Stanford, Berkeley, Alpino, MaltParser), the syntactic tree aligners Lingua-Align and the Dublin Sub-Tree Aligner.

Linguistic software

MultiTerm, WordSmith, Alchemist, Stockholm TreeAligner.

LEADERSHIP POSITIONS

2006-2007	Centre for Text Technology (CTexT) at the Potchefstroom campus of the North-West University:

- Project leader of the lexical database *ALEXANDER* (Afrikaans Lexicon and Annotated Database for Engineering and Research) (2006)
- Project leader of the Afrikaans WordNet (2007 – see "EDUCATION")

EDUCATION

Institutions:
- Rijksuniversiteit Groningen (University of Groningen) in Groningen, The Netherlands
- The former University of Port Elizabeth (presently the Nelson Mandela Metropolitan University)
- Eberhard Karls University in Tübingen, Germany

- Vrije Universiteit (Free University) in Amsterdam, The Netherlands
- Distance course in General Linguistics at the University of South Africa (2001)

2011   Attended European Summer School in Logic, Language and Information (ESSLLI) in Ljubljana, Slovenia.

2011   Attended a Machine Learning course at the University of Groningen, resulting in a paper with the title "Authorship attribution of weblog texts: A comparative study".

2010   Attended ESSLLI summer school in Copenhagen, Denmark.

2010   Attended the 4th Machine Translation Marathon at Dublin City University in Dublin, 2010.

2009   Attended ESSLLI summer school in Bordeaux, France.

2008-2013   **Doctor of Philosophy (Computational Linguistics)**

2013   Doctoral dissertation completed and accepted for public defense: *Complementary approaches to tree alignment: Combining statistical and rule-based methods.* In this work, the alignment of syntactic phrase-structures of translated texts is researched in order to create so-called *parallel treebanks*, of which the main application is syntax-based machine translation. Although the effect on the performance of such translation is found lacking, it is shown that various different novel approaches to the alignment problem can be successfully combined to create a high-quality resource.

2005-2007   **Master of Arts in Linguistics (Lexicology and Terminology)**

2007   Thesis completed: *Building a WordNet for Afrikaans: Preliminary research in the form of an inquiry into the feasibility and optimal methodology for the development of a wordnet database.* The work is a preliminary study about a proposed project at the North-West University in Potchefstroom, South Africa. It deals with the design and building of an Afrikaans wordnet – a semantic network stored as an electronic lexical database – for lexicographic and human language technological purposes. The project has been sponsored by the South African Department of Science and Technology.

2005-2006   Courses at the Vrije Universiteit (Free University) in Amsterdam, The Netherlands (names are in Dutch):
"Lexicografische en Terminologische Tools" (Lexicographic and Terminological Tools), "Bilinguale Lexicografie" (Bilingual Lexicography), "Werkcollege Lexicologie" (Seminar in Lexicology), "Werkcollege Terminologie" (Seminar in Terminology), "Lexicale Semantiek" (Lexical Semantics), "Computationele Lexicologie" (Computational Lexicology) (no examination), "Relationele Modellen" (Relational Models), "Vaktaal" (Technical Language).

2004-2005   **B.A. Honours in Applied Language Studies (Translation Studies)** (*cum laude*)

2004   Courses at the University of Port Elizabeth:
Introduction to General Linguistics, Applied Linguistics, Theory of Translation.
Courses at the Eberhard Karls University in Tübingen, Germany:
Computational Linguistics (Parsing and Programming in Java (Data Structures)), Advanced German, Spoken German.

2000-2003	**B.A. in Applied Language Studies** (*cum laude*)
2003	Courses: Afrikaans III (mainly Linguistics and Literature), German III, French III and Music Technology.
2001-2002	Courses: Afrikaans I and II, German I and II and French I and II. Additional Courses: Programming in C++, Programming in Java (Algorithmics and Data Structures, Web Technology, Mathematics (Algebra and Calculus), Information Systems and Computer Architecture.
2000	Courses: Afrikaans, Dutch, Introduction to University Practice and Computer Literacy.

PUBLICATIONS

Gideon Kotzé. 2006. *Building a WordNet for Afrikaans: Preliminary research in the form of an inquiry into the feasibility and optimal methodology for the development of a wordnet database*. Master's Thesis. Vrije Universiteit (Free University), Amsterdam.

Gideon Kotzé. 2008. Development of an Afrikaans wordnet: methodology and integration / Ontwikkeling van 'n Afrikaanse woordnet: metodologie en integrasie. In *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies: Human language technology for South African languages*. Special Issue 1, Volume 29, pp. 163-184.

Gideon Kotzé. 2011. Finding Statistically Motivated Features Influencing Subtree Alignment Performance. In *NODALIDA 2011 Conference Proceedings*, pp. 332-335. Riga, Latvia.

Gideon, Kotzé. 2011. Improving syntactic tree alignment through rule-based error correction. In *Proceedings of ESSLLI 2011 Student Session*. p. 122-127. Ljubljana, Slovenia.

Gideon Kotzé. 2011. Rule-induced error correction of aligned parallel treebanks. In *Proceedings of Corpus Linguistics*. p. 35-40. Saint-Petersburg, Russia.

Gideon Kotzé. 2012. Transformation-based tree-to-tree alignment. In *Computational Linguistics in the Netherlands Journal*. Vol. 2, pp. 71-96.

Gideon Kotzé. 2013. *Complementary Approaches to Tree Alignment: Combining Statistical and Rule-Based Methods*. PhD thesis. University of Groningen.

Gideon Kotzé, Vincent Vandeghinste, Scott Martens and Jörg Tiedemann. 2012. Large aligned treebanks for syntax-based machine translation. In *Proceedings of LREC 2012*. Istanbul, Turkey.

Jörg Tiedemann and Gideon Kotzé. 2009. A Discriminative Approach to Tree Alignment. In *Proceedings of the International Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography and Language Learning (in connection with RANLP'09),* pp. 33-39.

Jörg Tiedemann and Gideon Kotzé. 2009. Building a Large Machine-Aligned Parallel Treebank. In *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories (TLT'08)*, pp. 197-208, EDUCatt, Milano/Italy.

Vincent Vandeghinste, Scott Martens, Gideon Kotzé, Jörg Tiedemann, Joachim van den Bogaert, Koen De Smet, Frank Van Eynde, Gertjan van Noord. 2013. Parse and Corpus-based Machine Translation. In *Essential Speech and Language Technology for Dutch: resources, tools and applications*. Springer.

Vincent Vandeghinste, Scott Martens, Gideon Kotzé and Joachim van den Bogaert. 2011. PaCo-MT: Parse and Corpus-based Machine Translation. In *Proceedings of*

the 15th Annual Conference of the European Association for Machine Translation (EAMT 2011).

Talks and presentations
- Defence of my doctoral dissertation at the University of Groningen on June 24th 2013, resulting in the conferment of a PhD degree.
- Poster presentation at the Student Session of ESSLLI 2011 in Ljubljana, Slovenia. Title: Improving syntactic tree alignment through rule-based error correction. (published)
- Oral presentation at Corpus Linguistics 2011 in Saint Petersburg, Russia. Title: Rule-induced error correction of aligned parallel treebanks. (published)
- Poster presentation at NODALIDA 2011 in Riga, Latvia. Title: Finding Statistically Motivated Features Influencing Subtree Alignment Performance. (published)
- Talk at Computational Linguistics in the Netherlands (CLIN) 2011 in Ghent, Belgium. Title: Finding Statistically Motivated Features Influencing Subtree Alignment Performance.
- Invited in 2010 by Prof. Dr. Martin Volk as a colloquium speaker at the Institute for Computational Linguistics at the University of Zürich. The topic was the tree-to-tree alignment tool Lingua-Align developed by Jörg Tiedemann. I presented the published conference paper authored by Tiedemann at LREC 2010 in Malta.
- Oral presentation at TABU Dag 2010 in Groningen. Title: Supervised tree alignment for syntax-based machine translation.
- Poster presentation at Computational Linguistics in the Netherlands (CLIN) 2010 in Utrecht, The Netherlands. Title: Discriminative parallel treebank alignment for syntax-based machine translation.
- Oral presentation presented at TLT'09 in Milan, Italy. Published in proceedings and co-authored with Jörg Tiedemann. Title: Building a Large Machine-Aligned Parallel Treebank. (published)
- Poster presentation at TABU Dag 2009 in Groningen. Title: Training a statistical parser for parsing French for use in syntax-based machine translation.
- Oral presentation presented at Computational Linguistics in the Netherlands (CLIN) 2009. Title: Automatic filtering of parallel corpora for improving alignment accuracy.
- Oral presentation presented at ALASA conference (International Conference of the African Language Association of Southern Africa) in Port Elizabeth, South Africa, 9-11 July 2007. Title: *Development of an Afrikaans wordnet: methodology and integration.*
- Oral presentation at LSSA/SAALA/SAALT joint conference (Linguistic Society of Southern Africa (LSSA), the South African Applied Linguistics Association (SAALA) and the South African Association for Language Teaching (SAALT)) in Potchefstroom, South Africa, 4-6 July 2007. Title (in Afrikaans): *Ontwikkeling van 'n Afrikaanse woordnet: metodologie en integrasie* (Development of an Afrikaans wordnet: methodology and integration)

Scholarships, awards, funding and others

- Obtained funding for partaking in a machine translation project, PaCo-MT, from 2008-2011 as part of my PhD program at the University of Groningen (2008-2013). After completion of the project in 2011, the University funded the rest of my research activities up to the end of 2012.
- ZASM (Zuid-Afrikaanse Spoorwegmaatschappij) scholarship for Masters study in Amsterdam, The Netherlands (2005-2006)
- Rutgers scholarship from the Vrije Universiteit's Department of Arts (Letteren) for Masters study in Amsterdam, The Netherlands (2005-2006)
- Project funds received at the South African Department of Science and Technology for the construction of an Afrikaans wordnet (see "EDUCATION") at the Centre for Text Technology (CTexT) at the Potchefstroom campus of the North-West University, South Africa (2006)
- Member of the *Landesstiftung Baden-Württemberg* in Germany (a foundation enabling scholars, students and young working people to study in Baden-Württemberg, Germany) (2004)
- Chosen as exchange student (representing the former University of Port Elizabeth) for the summer semester of 2004 at the University of Tübingen, Germany, with the simultaneous conferment of the Baden-Württemberg scholarship for the exchange programme (2004)
- Postgraduate scholarship at the University of Port Elizabeth (2004)
- Dean's scholarship at the University of Port Elizabeth (2004)
- Sponsored as participant in a French course and cultural programme at the University of Réunion, Réunion (France) (2003)
- Member of the Golden Key International Honour Society (2003)
- DAAD (*Deutscher Akademischer Austausch Dienst*) scholarship for German study in South Africa (2003)
- Merit bursaries at the University of Port Elizabeth (2000-2003)
- Dux pupil: Matric 1999 (best position in final year at high school)

INTERESTS AND HOBBIES:

I am interested in all the various ways in which we can describe and model language, how languages are the same and how they differ from each other, all the various useful things which we can do with all this knowledge, and what it tells us about the human experience. More specifically, I am interested in disciplines such as corpus linguistics, lexicography, machine translation, manual and automatic tree alignment, language modeling, syntactic parsing, grammar acquisition and automated or semi-automated lexical database construction.

Many of these fields of expertise also have direct relevance to the industry as companies become more global and web-based. Intelligent processing of language online is becoming an indispensable requirement. As far as computational linguistics and natural language processing are concerned, the boundaries between the academics and the industry are becoming fuzzier. I find this change of landscape very intriguing and am excited about the prospects of being part of it.

In my free time, I like to play the piano, video games, chess and tennis, watch fiction, documentaries and sports, do programming, write creatively, travel, read

fiction, spend time in nature, and read and discuss philosophy, natural science and technology. I also have a general interest in human culture, behaviour and current affairs, and often dream about a bright future featuring an improved standard of life, exciting new technologies, space exploration and increased human understanding, both of the world around us and of each other.