University of Zurich[UZH]

Master's thesis
for the degree of
Master of Arts
presented to the Faculty of Arts and Social Sciences
of the University of Zurich

Grasping the Nettle: Neural Entity Recognition for Scientific and Vernacular Plant Names

Author: Isabel Meraner
Student ID: 16-741-514

Examiner: Prof. Martin Volk

Supervisor: Nora Hollenstein

Institute of Computational Linguistics

Submission date: 01.03.2019

Abstract

The recognition of taxonomic entities plays a key role for natural language processing and understanding in botanical contexts. Automating knowledge and entity extraction is hence a core endeavor for populating and enriching existing knowledge bases and for digitally interlinking scientific and ethnobotanical plant knowledge. In this work, we present a semi-supervised approach for the automatic recognition of plant names in their Latin and vernacular variants in German and English across different text genres. For this purpose, we compiled four corpora differing in terms of formality (scientific vs. non-scientific), writing style (blog articles vs. introductory Wikipedia sections), and publication period (historical vs. recent). Our proposed pipeline includes linguistic preprocessing based on sentence splitting, tokenization, and part-of-speech tagging. Additionally, we use dictionary-based annotations to automatically label the corpus data and create a low-effort silver standard. The dictionary lookups rely on large language-specific gazetteers for a total of nine hierarchical scientific and vernacular entity labels collected from several botanical resources. Subsequently, we train a state-of-the-art named entity recognition (NER) system based on a bidirectional long-short-term-memory architecture [Hochreiter and Schmidhuber, 1997] followed by a conditional random field layer (bi-LSTM-CRF) [Lample et al., 2016]. To exploit token-level and character-level contextual information from the silver-labeled datasets, we integrate the 300-dimensional pretrained *FastText* word embeddings [Grave et al., 2018] and re-train character-level word representations on the input data [Lample et al., 2016]. In total, we generate eight neural models per language and dataset. An evaluation of the entity tagger shows F_1-scores of >86% on both manually and automatically annotated test sets for the combined English dataset. For German, we report a final F_1-score of >94% on both annotation types. We discuss the insights gained from adopting several dataset and language-specific parameter combinations from a single and cross-dataset evaluation perspective. Finally, we disambiguate the entity candidates proposed by the tagging system and link them to an international botanical reference database using a lookup table for the vernacular names. Our approach emphasizes the potential of fine-grained, domain-specific entity labels and low-effort data models trained on automatically labeled corpus data to explore and computationally process lower-resourced fields and genres for knowledge preservation purposes.

Zusammenfassung

Die hier vorliegende Masterarbeit befasst sich mit der automatischen Erkennung, Klassifizierung und Verlinkung von wissenschaftlichen und volkstümlichen Pflanzennamen in Texten. Zu diesem Zweck wurden insgesamt vier Korpora für Deutsch und Englisch mit Texten unterschiedlicher Genres, Schreibstile und Epochen erstellt, die anschließend mithilfe eines Wörterbuch-basierten Annotationssystems automatisch annotiert wurden. Letzteres beruht auf umfangreichen Namenslisten für lateinische, deutsche und englische Pflanzennamen, die für diese Arbeit aus diversen botanischen Ressourcen zusammengeführt wurden. Anschließend wurden verschiedene neuronale Modelle mithilfe eines bidirektionalen LSTM-CRFs für beide Sprachen trainiert und ausgewertet. Die vom finalen System vorgeschlagenen Pflanzennamen wurden abschließend mithilfe einer botanischen Referenzdatenbank disambiguiert und verlinkt. Diese Arbeit trägt somit dazu bei, multilinguales, botanisches sowie biodiversitätsbezogenes Wissen zu erschließen und Pflanzen-Volksnamen in verschiedenen Textsorten automatisch zu erkennen und zu disambiguieren.

Acknowledgement

First, I would like to express my sincere gratitude to Prof. Dr. Martin Volk who supported my idea to apply multilingual text analysis to the field of botany and to biodiversity science from the very beginning. I am also truly grateful to my mentor, Nora Hollenstein, for her technical expertise, her endless patience and her motivational and inspirational support throughout my thesis project. In addition, I would like to thank Phillip Ströbel for his valuable instructions on how to digitize the historical works during the initial phase of this project and Mathias Müller for helping me with our GPU server. Very many thanks go to Donat Agosti and Guido Sautter from *Plazi* in Bern, for sharing their digitized botanical works and technical experience in the field of biodiversity informatics. A big thank you also goes to Vienna, especially to Heimo Rainer from the *Museum of Natural History Vienna* and Megumi Kurobe, and the inspiring talks and refreshing moments of knowledge sharing we had over the past months.

Many thanks also go to the numerous botanists and institutions for sharing their data and knowledge, especially to Reto Nyffeler from the *Botanical Garden Zurich*, Wilfred Gerritsen from the *Catalogue of Life*, Paul Hemetsberger from *dict.cc*, Florian Bärtschi from the *Botanical Garden Basel*, Michael Jutzi from *infoflora.ch*, Thomas Wilhalm from the *South Tyrolean Museum of Natural History Bolzano*, Felix Brüngger from *igarten.ch*, Helmut Knüpffer from *IPK Gatersleben*, Stefan Imhof from *University of Marburg* and Werner Arnold from *awl.ch*.

Last but not least, I would also like to thank from the bottom of my heart my fellow student Alessia for sharing her joy, inspiration, and moments of crisis with me, and, of course, my parents and Michael, for motivating and encouraging me throughout my studies in Multilingual Text Analysis at the University of Zurich. And finally, a big thank you to Albin, Alessia, Caroline and Thomas for proofreading my thesis and providing me with valuable feedback. Without the help of every single one of you, this personal achievement would not have been possible.

Contents

List of Figures

List of Tables

List of Acronyms

API	Application Programming Interface
CBOW	Continuous Bag Of Words
CRF	Conditional Random Field
EL	Entity Linking
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
ID	Identifier
JSON	JavaScript Object Notation
LOD	Linked Open Data
LSTM	Long-Short-Term-Memory
ME	Maximum Entropy
MWE	Multiword Expression
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OCR	Optical Character Recognition
PDF	Portable Document Format
PoS	Part-of-Speech
SKOS	Simple Knowledge Organization System
SGD	Single Gradient Descent
UTF-8	Unicode Transformation Format (8-bit)
WSD	Word Sense Disambiguation
XML	eXtensible Markup Language

* * *

"There's rosemary, that's for remembrance; pray, love, remember: and there is pansies; that's for thoughts. There's fennel for you, and columbines. There's rue for you; and here's some for me; we may call it herb-grace o'Sundays: O you must wear your rue with a difference. There's a daisy. I would give you some violets, but they withered all when my father died. They say he made a good end."

* * *

— William Shakespeare, *Hamlet* (Act 4, Scene 5)

1 Introduction

1.1 Motivation

International knowledge about plants, their value for human alimentation, and their pharmacological application in folk medicine and herbalism is encoded in countless languages. This wealth of traditional expertise concerning the structure, nutritional value, chemical constituents and established plant-based treatments is of inestimable value for our society. Despite extensive digitization efforts at present, this traditional, plant-related knowledge is not always accessible in a digital, human-readable and machine-readable format. The identification of scientific and vernacular plant names in multilingual text material can be a way to approach the task of information extraction in lower-resourced and under-represented fields, such as ethnobotany and local biodiversity research. Aggregating large amounts of available information into publicly accessible, sustainable and internationally standardized formats and databases still constitutes a great challenge in many fields of research. The adoption of Linked Open Data (LOD) for the purpose of collecting, structuring, sharing, and connecting precious pieces of knowledge for under-resourced domains has a great potential not only for linguistics, but also for botany, biodiversity management, and related fields [Bizer et al., 2008; Lehmann et al., 2015; Chiarcos et al., 2011; Minami et al., 2013].

Modern botanical literature is commonly held in English and almost exclusively employs scientific Latin plant names. Integrating taxonomic entities (taxa) on a vernacular level can be particularly beneficial when extracting and processing plant knowledge from historical, heterogeneous texts or non-scientific text genres [Sharma et al., 2017]. Additionally, the adoption of a multilingual perspective based on the languages German and English can reveal fascinating linguistic properties of the text passages encoding plant knowledge and the shape, quality, and composition of botanical entities in natural languages. To automatically build or enrich multilingual knowledge bases for such specific domains, it is essential to aggregate and interlink vernacular or synonymous alternative names to their associated and currently accepted Latin identifiers [Seideh et al., 2016b]. By this means, compu-

tational methods contribute to approach the so-called botanical "names problem" [Boyle et al., 2013; Patterson et al., 2010] and to disambiguate vernacular, synonymous names, and outdated Latin name variants. Our project has been motivated by interdisciplinary efforts to apply linguistic and taxonomic expertise in combination with state-of-the-art neural named entity recognition (NER) methods to the fields of biodiversity, (ethno-)botany, and folk medicine. With our approach, we aim at making such domains more accessible and, last but not least, not only to preserve knowledge in a machine-readable way, but also to safeguard centuries-old autochthonous plant names, associated treatments and traditions.

1.2 Task Description and Outline

In this work, we present a semi-supervised approach for scientific and vernacular plant name recognition and classification across different text genres for German and English.[1] Table 1 summarizes our overall approach and the single stages of the project. First, we collect language-specific gazetteers (name lists) for English, German and Latin plant names (stage (1)). We compile four distinct corpora (stage (2)) for the languages English and German containing different text genres and writing styles (Wikipedia articles, mountaineering reports, historical and modern botanical literature, blog articles). To guarantee a high concentration of botanical entities, we filter the sentences for at least one present vernacular or scientific entity mention. We then apply linguistic preprocessing (stage (3)) before automatically annotating the data in stage (4) using a tailored dictionary-based system. In stage (5), we train a state-of-the-art neural named entity recognition (NER) system [Lample et al., 2016] on these datasets exploiting token-level and character-level distributional information of natural language [Mikolov et al., 2013; Ling et al., 2015; Huang et al., 2015]. While many common NER approaches focus on the recognition of predefined entity labels such as organizations (ORG), locations (LOC), and persons (PER), we propose a fine-grained label set for nine domain-specific entity classes. In total, we use two labels for vernacular names and seven labels for Latin names to detect taxonomic entities in German and English texts. Instead of focusing on the evaluation and application on common benchmark corpora for NER such as the CoNLL-2003 dataset [Tjong Kim Sang and De Meulder, 2003], we apply and evaluate our models using the carefully selected domain-specific datasets in a single and cross-corpus evaluation setting (stage (6)). To improve the quality of the training material and the resulting model performance after an initial evaluation round,

[1] We made the annotated corpora and scripts available to the research community on GitHub: `https://github.com/IsabelMeraner/BotanicalNER`.

we re-annotate the data in stage (7) using pattern-based corrections and apply these cleaned datasets for re-training and evaluating the neural models. In stage (8), we use the *Catalogue of Life* (CoL) [Roskov et al., 2018] to link and disambiguate the entity candidates proposed by the tagger. To provide an end-to-end entity recognition and linking service for extracting botanical names from raw, unstructured text, we implement a companion web-interface in stage (9). We argue that state-of-the-art natural language processing methods can be successfully applied to multifaceted text genres and styles in order to detect and disambiguate entities on a scientific and vernacular level, thus contributing to the exploration, interlinkage, and storage of international botanical knowledge.

Project Stage	**Subtask**	**Resources/Methods/Tools**
(1)	Creation of gazetteers by aggregating (and, if necessary, digitizing) several resources containing scientific and vernacular plant names for German and English.	botanical/biodiversity databases, private botanists and institutions, historical botanical literature
(2)	Creation of four distinct training corpora from manifold resources including various text genres and writing styles.	Wikipedia articles, Text+Berg corpus [Volk et al., 2010], blog articles and botanical literature
(3)	Linguistic preprocessing of datasets including tokenization, sentence segmentation, POS-tagging and lemmatization.	NLTK [Loper and Bird, 2002], spaCy [Honnibal and Montani, 2017]
(4)	Dictionary-based automatic annotation of the training data using the gazetteers created in the first stage. Adoption of nine distinct entity labels for data annotation: 2 for German/English vs. 7 for Latin.	Annotations in IOB scheme [Sang and Veenstra, 1999] Tailored dictionary-based NER-tagger
(5)	Training of multiple models for German and English using a bidirectional LSTM-CRF architecture.	bi-LSTM-CRF tagger by Lample et al. [2016]
(6)	Single and cross-dataset evaluation of German and English models using different datasets and parameter combinations.	Metrics: precision, recall, F1-score
(7)	Re-annotation of training data using pattern-based corrections and subsequent re-training and evaluation of the models.	bi-LSTM-CRF [Lample et al., 2016]
(8)	Disambiguation and linking of identified candidate entities in a botanical reference database.	Catalogue of Life [Roskov et al., 2018]
(9)	Implementation and visualization on web-interface: tokenization of input, tagging and linking of candidates.	Bootstrap [Getbootstrap.com, 2015], Flask [Grinberg, 2014]

Table 1: Stages of project and associated subtasks, resources, methods, and tools.

1.3 Research Interests and Contributions

In this project, we apply state-of-the-art neural methods with the central aim of identifying not only scientific, but also vernacular plant names in manifold text genres for German and English. By integrating a dictionary-based annotation system in our pipeline (project stage (2)), we avoid time-consuming, manual human annotation. Nonetheless, we are able to guarantee large, high-quality silver standard datasets for training and testing the models (see Section 4.1). Another central concern is to explore the multilingual applicability of state-of-the-art named entity recognition (NER) methods to under-explored text genres, such as mountaineering reports or botanical literature. Moreover, our approach emphasizes the potential of domain-specific fine-grained entity labels and low-effort data models trained on automatically annotated material to explore and computationally process such lower-resourced fields and genres. It is in our best interest to exemplify that botanical and linguistic expertise in combination with state-of-the-art natural language processing (NLP) methods and technologies can develop a symbiosis at an interdisciplinary level to promote biodiversity science, preserve knowledge and revive the interest in our world's botanical heritage.

The research questions of central interest in this project are:

1. How well does the state-of-the-art bidirectional LSTM-CRF architecture for named entity recognition perform on domain-specific scientific and non-scientific text genres?

2. How does the performance vary regarding morphologically simple versus rich languages, namely English versus German? How do linguistic phenomena such as compounding influence the performance of the neural NER models?

3. How well does the tagger perform regarding different classes of entities and different taxonomic levels for scientific and vernacular names?

4. Can neural models trained on large, low-effort datasets such as Wikipedia be applied to robustly identify botanical entities in datasets from low-resourced domains, where only limited data resources might be available for training?

5. Can entity linking be applied for the semantic disambiguation of both scientific and vernacular plant names?

To the best of our knowledge, this work is the first endeavor to apply computational methods for the recognition of both scientific (Latin) and multilingual vernacular plant names (English and German) in text. Hence, this project does not only con-

tribute to the enhancement of current NER systems in the biodiversity and botany domain, it also addresses the challenging task of automating knowledge extraction in domain-specific, scientific and non-scientific text material. By disambiguating and interlinking scientific and vernacular names to international taxonomic reference databases, our approach provides the possibility to extend and enrich such biodiversity knowledge bases with multilingual, regional vernacular name variants and associated knowledge. Moreover, this project can contribute to open up the fields of botany and biodiversity management to a broader public by integrating knowledge on a vernacular level. For instance, we are in touch with members from biodiversity projects and databases, such as the *Catalogue of Life* (CoL) [Roskov et al., 2018], to improve the query functions of the web-service. In summary, we sustain that the integration of vernacular plant names and associated knowledge into open-source knowledge bases holds promising opportunities for safeguarding traditional, ethno-cultural heritage related to plants [Seideh et al., 2016b; Sharma et al., 2017], which is frequently encoded in material from poorly explored domains, such as ethnobotany and folk medicine.

1.4 Thesis Structure

In Chapter 2, we embed the project in its theoretical background and give an overview on botanic naming traditions. We explain several methods that have been used to approach the task of named entity recognition (NER). We present notable task-related projects in the field of botany and biodiversity informatics before introducing the most important botanical knowledge bases and entity linking approaches. Chapter 3 describes the data resources and tools that we used to aggregate the language-specific gazetteers and training corpora. We explain the preprocessing steps involved during data preparation and the subsequent dictionary-based annotation of the corpora. Moreover, we illustrate the annotation guidelines for the additional creation of a gold standard to evaluate the bi-LSTM-CRF models. At last, we sketch our approach to disambiguate and link the entity candidates to a botanical reference database. In Chapter 4, we present a detailed evaluation of our models in a single-corpus and cross-corpus evaluation setting. We conduct an error analysis to explore potential pitfalls and sources of error. Chapter 5 focuses on the outcomes and results gained from the final project stage concerned with the disambiguation and linking of the entity candidates. We discuss our overall insights from this project and return to the contributions and possible future applications of this work in Chapter 6 and conclude the thesis in Chapter 7.

2 Related Work

2.1 The Names of Plants

The endeavor of naming organisms on a scientific and vernacular level has a long tradition and is indispensable in order to explicitly refer to an individual. Latin plant names and their typical suffixes are additionally designed to classify multiple instances into genera, families, subfamilies and other hierarchical classes. Conversely, *vernacular names* (also referred to as *common* or *trivial plant names*) are prone to linguistic changes over time. They are also highly dependant on language, culture, and, of course, flora and fauna of a specific region. According to Roskov et al. [2018], less than a fifth of the organisms on this world has so far been identified, named, and catalogued. Over time, this condition has triggered manifold local and global taxonomic initiatives and traditions with the central goal of naming and documenting these testimonies of the world's biodiversity. Many vernacular names are synonymous, ambiguous, or misleading such as *corn* (Zea mays), which is a synonym of *maize* and is also used in British English to refer to different types of cereals, e.g. *wheat* and *rye* (in Scotland) [Bareja, 2010]. The well-known *dandelion* (Taraxacum officinale), for instance, has more than 500 vernacular names in standard German and local dialects, such as "Löwenzahn", "Pfaffendistel", "Pusteblume", "Milchschreck", "Kuhblume" or "Ackerzichorie" [Helmut, 1986], just to mention a few of them. The need for an international classification system based on formal consensus was first explored by Aristotle and later extended in the polynomial system, which resulted in excessively long descriptive names [Bareja, 2010; Hogan and Taub, 2011]. With Charles Linnaeus, the so-called father of modern taxonomy, a simplified naming tradition entered the field, today known as "binomial nomenclature" using a capitalized generic name followed by a lowercase species epithet [Knapp, 2000]. For the *European goldenrod*, the binomial nomenclature results in Solidago$_{[GENUS]}$ virgaurea$_{[EPITHET]}$. Nonetheless, several parallel and often incompatible naming conventions such as the *Phylogenetic Nomenclature* [Cantino et al., 1999] exist at present. Over the past, the effort of individual botanists to name local plants, resulted in synonymous and alternative names beyond number. Among taxonomists

and botanists this unfortunate situation is also referred to as the "names problem" [Boyle et al., 2013; Patterson et al., 2010]. The largest challenge is *homotypic synonymy*, describing the existence of multiple names for one species, a situation that prevents new information and research on the same taxon being indexed to the same taxonomic instance [Boyle et al., 2013; Patterson et al., 2010]. *Heterotypic synonymy*, on the other hand, occurs when different taxa are treated as being the same identical species [Patterson et al., 2010]. In addition to that, we might encounter lexical variants, misspellings and the phenomenon of *homonymy*, which is the case when the same name is used to refer to different distinct taxa [Boyle et al., 2013]. According to Boyle et al. [2013], 5% to 20% of all published names are synonymous, which makes taxonomic standardization a cumbersome and error-prone task. The following names described by different botanists, for instance, all refer to the fragrant herb 'lemon verbena' (*homotypic synonymy*): *Aloysia sleumeri*, *Aloysia citrodora*, *Aloysia triphylla*, *Lippia triphylla*, *Lippia citrodora*, *Verbena triphylla*, *Verbena citrodora*, *Verbena fragrans*, *Zappania citrodora*. Among these, the currently accepted scientific name according to the international reference database *Catalogue of Life* (CoL) is *Aloysia citrodora* [Roskov et al., 2018].

In this project, we address this "names problem" [Boyle et al., 2013; Patterson et al., 2010] from a pragmatic perspective: How do these "names" look like in natural language, in which contexts do they occur, what shape and orthographic properties do they have, and how can we reliably identify and disambiguate them? To model such shape-based and context-related evidence, we use character-level and token-level distributional information [Mikolov et al., 2013] to represent our input data. We then apply state-of-the-art neural methods for named entity recognition (NER) [Lample et al., 2016] to train multiple German and English models with the goal to automatically identify individual tokens or sequences of tokens referring to either vernacular or scientific names. Finally, we explore possible ways to disambiguate these name occurrences in natural language and to interlink them to unique entries in botanical knowledge bases. We hope that the comprehensive integration of vernacular names into international databases contributes to open up the domain of botany to a broader public that might be only familiar with vernacular names.

2.2 Named Entity Recognition

Named entity recognition (NER) is an essential subtask of information extraction and natural language processing and understanding with the central concern of finding sequences of tokens that constitute an information unit referring to a named

entity [Nadeau and Sekine, 2007; Jurafsky and Martin, 2018]. Besides different recent neural architectures, the NER task has also been approached using various rule-based, dictionary-based or first generation machine-learning techniques [Tjong Kim Sang and De Meulder, 2003], which we will briefly introduce and compare in this chapter. Due to the two-sided nature of the task, NER is often also referred to as *NERC* ("named entity recognition and classification") [Nadeau and Sekine, 2007]. The initial stage of the NER task focuses on the detection of named entities in text whereas the second substep aims at classifying the detected instances into an according entity class such as "person", "organization" or "location" [Nadeau and Sekine, 2007]. Depending on the target domain and the specific application, these entity classes or types can be coarse-grained or fine-grained and involve from three basic categories (person, organization, location) up to 200 distinct classes [Nothman et al., 2013; Ekbal et al., 2010; Ratinov and Roth, 2009; Nadeau and Sekine, 2007]. The main challenges of the NER task are *type ambiguity* (an entity might refer to different possible referents in the real world) and the *ambiguity of segmentation* [Jurafsky and Martin, 2018]. Commonly, NER approaches rely on word-by-word sequence labeling to assign IOB tags marking the boundaries (Inside, Outside, Beginning) and the entity type of the mention [Sang and Veenstra, 1999; Ratinov and Roth, 2009; Jurafsky and Martin, 2018].

In this project, we apply the standard neural algorithm for NER, namely a bidirectional LSTM (originally coined by Hochreiter and Schmidhuber [1997]) with a final CRF-layer. More specifically, we work with the bi-LSTM-CRF architecture proposed by Lample et al. [2016] using character-level and token-level word representations as input information. This neural approach to NER is both domain-independent and language-independent and does not resort to hand-engineered features or knowledge resources during training [Lample et al., 2016; Ma and Hovy, 2016]. When processing morphologically rich languages such as German, character-level word representations are particularly beneficial and refine the models with regard to language, domain, and text genre [Lample et al., 2016; Ling et al., 2015]. Similar to weakly and semi-supervised approaches, we use large gazetteers of plant names for the initial dictionary-based annotation of our training material, which is then used as high-quality input material for the bi-LSTM-CRF.

2.2.1 Rule-Based versus Dictionary-Based Approaches

Early named entity recognition (NER) approaches make use of handwritten rules, algorithms, or patterns to detect named entities in running text [Nadeau and Sekine, 2007]. Due to the time-consuming task of creating and adapting such rules and the

resulting non-applicability to new domains and text genres, rule-based systems have largely been replaced by machine learning techniques especially in academic research [Chiticariu et al., 2013]. In contrast, botanical rule-based systems such as *FAT* (Find All Taxon Names) [Sautter et al., 2006] dealing with the detection of Linnean plant names and systematic Latin suffixes can achieve satisfying results. Especially with regard to entities following a predefined orthographic pattern or limited sets of entities in a specific domain, rule-based or hybrid systems work reliably in real-world applications [Chiticariu et al., 2013].

As opposed to rule-based systems, dictionary-based NER systems heavily rely on extensive gazetteers (also called "lists", "lexicons" or "dictionaries" [Nadeau and Sekine, 2007]) that need to be constantly expanded as soon as new entities are coined in a language. Such systems perform lookups in domain-specific dictionaries to detect and annotate the entities of interest [Shinzato et al., 2006]. To reduce the effort and resources needed for manually annotating large corpora for further training tasks, such dictionary-based methods can be a promising approach for specific domains [Nadeau and Sekine, 2007; Shinzato et al., 2006]. According to Basaldella et al. [2017], hybrid approaches can be beneficial for specific domains such as biomedicine: Such approaches usually combine a dictionary-based system for entity candidate generation with machine learning methods to filter and select relevant candidates.

Similarly, in our project we are using a dictionary-based NER system to generate automatically annotated, high-quality training material. Even though we did not fully automatically construct the final dictionaries as demonstrated by Shinzato et al. [2006], we managed to avoid time-consuming manual annotation and to generate a large silver-labeled corpus for the subsequent neural training.

2.2.2 Weakly and Semi-Supervised Learning Approaches

Other information extraction (IE) approaches use bootstrapping techniques to extract semantic resources from text in an iterative setting [Curran et al., 2007]. Iterative learning settings exploiting contextual patterns and morphological evidence achieve promising performance, especially with regard to language-independent applications [Cucerzan and Yarowsky, 1999]. To avoid time-consuming and knowledge-intense manual annotation, weakly and semi-supervised approaches use seed lists of named entities during the annotation stage [Grave, 2014]. The resulting automatically labeled positive examples can then be used for supervised training. Fully unsupervised approaches for information extraction, on the contrary, do not rely on

feature engineering and labeled examples or seed terms but apply recursive extraction methods to retrieve new entities, relations or attributes [Etzioni et al., 2005].

Our method is also inspired by weakly and semi-supervised approaches: Instead of using seed lists of named entities for annotation [Grave, 2014], we integrate a dictionary-based annotation system into our pipeline. We use the resulting automatically labeled positive examples to train a bi-LSTM-CRF [Lample et al., 2016] in a supervised learning setting. In this way, we avoid devoting excessive time and resources to expensive expert-knowledge required for manual annotations.

2.2.3 Supervised Machine Learning Approaches

Prior to the neural generation of named entity recognition (NER) systems, the task has been approached using several supervised learning techniques. The predominant supervised learning methods include maximum entropy (ME) [Bender et al., 2003], hidden or conditional Markov models (HMM or CMM) [Morwal et al., 2012; Jansche, 2002], support vector machines (SVM) [Ekbal and Bandyopadhyay, 2010], decision trees [Paliouras et al., 2000], and conditional random fields (CRF) [Sato et al., 2017; Tjong Kim Sang and De Meulder, 2003]. Usually, in such learning settings all input observations are associated with a correct label in a large training corpus in order to learn and extract features given some contextual information during training [Jurafsky and Martin, 2018]. This supervised machine learning setting associating an expected output label $\mathbf{y}$ to each input observation $\mathbf{x}$ also applies to the neural generation of algorithms tackling the NER task, on which we will focus in the next subsection. In our approach, the expected output label $\mathbf{y}$ corresponds to the silver standard entity label that has been automatically generated during the dictionary-based annotation system.

2.2.4 State-Of-The-Art Neural Named Entity Recognition

Neural architectures using distributional embedding information as input have become increasingly important to efficiently solve sequence tagging tasks in natural language processing [Huang et al., 2015; Ma and Hovy, 2016]. Early neural approaches for named entity recognition (NER) include voted perceptrons [Carreras et al., 2003] and recurrent neural networks [Hammerton, 2003]. The network architecture based on the "long short-term memory" (LSTM) method coined by Hochreiter and Schmidhuber [1997] has proven to work particularly well on a series of natural language processing tasks, including NER. This network architecture repre-

sents a powerful and more complex variant of conventional recurrent neural networks (RNN). Instead of the commonly used hidden layer updates, the LSTMs make use of "purpose-built memory cells" [Huang et al., 2015] in combination with input, output and forget gates to remove information that is no longer needed and, in addition, to keep information that might be useful for subsequent decisions [Hochreiter and Schmidhuber, 1997; Huang et al., 2015]. Hence, non-local dependencies and information encoded in distant context can be identified and modeled successfully. In particular for named entity recognition, bidirectional LSTM architectures in combination with a conditional random field (CRF) layer achieve state-of-the-art performance in a supervised learning setting [Huang et al., 2015; Lample et al., 2016; Strauss et al., 2016]. Neural learning methods use word embeddings to represent the input data and to embed each observation in its left (l) and right (r) context in the bi-LSTM encoder (see Figure 1). Consequently, this bidirectional network architecture allows us to make use of past and future features encoded in the input representation [Huang et al., 2015]. On top of the bi-LSTM output, a CRF layer combines past input features from the LSTM-layers and tag information on the sentence level [Huang et al., 2015; Lample et al., 2016].

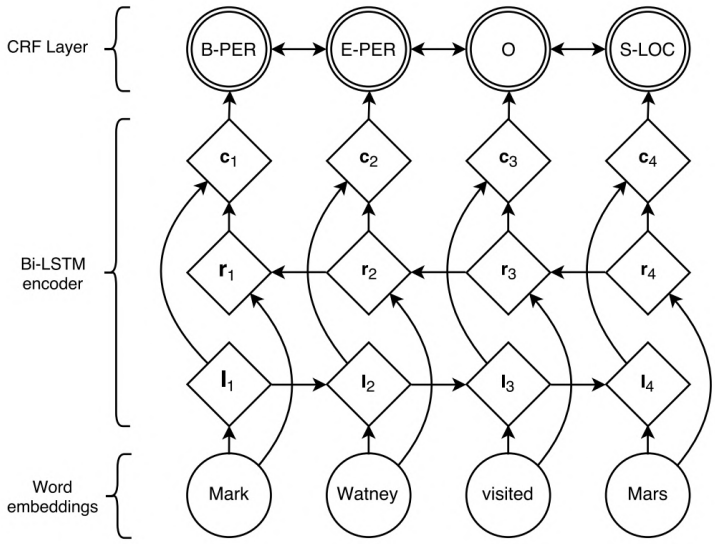

Figure 1: Bidirectional LSTM-CRF architecture from Lample et al. [2016].

In summary, such bidirectional LSTM-CRF models making use of past and future contextual input information can achieve state-of-the-art performance not only for the NER task, but also for part-of-speech-tagging and chunking tasks [Huang et al., 2015]. Thanks to past and future global contextual information from the input representation, any local labeling decision relies on non-local contextual evidence. During training, the neural system tries to estimate the expected outcome as it is

labeled in the annotated corpus. While a loss function models the gap between the prediction and the true output, a stochastic gradient descent optimization algorithm tries to find patterns and parameters in the data minimizing the loss function and, at the same time, maximizing the probability of a correct prediction [Huang et al., 2015; Jurafsky and Martin, 2018]. Nonetheless, it should be mentioned that such techniques are prone to overfitting and often fail to generalize well to unseen data. This is because the network tends to memorize the input training data and the shape and quality of the labeled entities [Hinton et al., 2012; Srivastava et al., 2014].

In our experiments, we explore several parameters of the bi-LSTM-CRF architecture and the resulting impact on the model performance for the task of identifying and correctly classifying botanical entities in multiple text genres. To assess the tendency to overfitting, we conduct experiments in a cross-corpus evaluation setting and test the model performance on unseen data from different genres (see Section 4.3) and on a held-out test set containing articles about fungi (see Section 4.4).

2.3 Botanical Databases and Entity Recognition

The existence of alternative name variants poses considerable challenges to the task of automatically identifying taxonomic entities in botanical knowledge-extraction [Boyle et al., 2013]. Multiple international standardization projects are devoted to address this issue by providing comprehensive check lists compiled from international taxonomic databases and resources.

2.3.1 Botanical Knowledge Bases

The *Catalogue of Life* (CoL) is the most comprehensive and authoritative online database with approximately 1.8 million catalogued organism species, including animals, plants and fungi [Roskov et al., 2018]. Not only does this institution provide an accurate, accessible, comprehensive, and international web service for private botanists and researchers, it also represents a platform for sharing and validating species names to document global biodiversity. Besides multiple global biodiversity projects such as the *Global Biodiversity Information Facility* (GBIF) [GBIF, 2018], the *Encyclopedia of Life* (EoL) [Hogan and Taub, 2011] or the *International Union for Conservation of Nature's* (IUCN) red list of threatened species [IUCN, 2018], totally 168 international databases contribute to this resource. Apart from proposing an internationally accepted scientific name for each species, it also lists outdated variants and, especially for English, associated vernacular names. The

International Plant Name Index (IPNI) provides bibliographical information about species from the seed plants, ferns, and lycophytes (spore-bearing vascular plants) [IPNI, 2012]. However, it does neither highlight currently accepted names nor does it list synonyms, vernacular names or spelling variants. The *Encyclopedia of Life* (EOL) [Hogan and Taub, 2011] gathers accessible knowledge (e.g. information about the habitat, geographic distribution or available images and publications) about any form of life on this planet and uses the annual *CoL*-checklists as a taxonomic data backbone [Roskov et al., 2018; Hogan and Taub, 2011]. The *Multilingual Multiscript Plant Name Database* (MMPND) [Porcher, 1999] is a distributed plant database containing scientific and non-scientific names of plants and fungi in 70 languages and 25 scripts. Regarding German, the MMPND contains only a few vernacular names for the genus *Brassica* ('cabbage'). The *Germplasm Resources Information Network* (GRIN) taxonomy project [Wiersema, 2018] includes 46,000 plant species with a strong focus on vascular plants of economic relevance. The *Global Biodiversity Information Facility* (GBIF) is a global network of participating countries and institutions hosting geo-referenced international biodiversity data, which can be used to assess the distribution of specimen records [GBIF, 2018]. The *uBio-project* [Norton et al., 2018] aims at providing a comprehensive catalogue to find information on living and once-living organisms. In the *NameBank* project, Patterson et al. [2006] focus on the importance of taxonomic indexing to manage data and research related to biology in the era of big data. In particular, taxonomic indexing enables the reconciliation of synonymous alternative names and the disambiguation of identical species names used for distinct taxa.

We adopt the latest annual *CoL* checklist [Roskov et al., 2018] to automatically extract botanical entities on different taxonomic levels. Especially with regard to Latin names and English vernacular names, this resource was of inestimable value for the creation of our gazetteers. For the sake of completeness, we also retrieve the available scientific and vernacular entities from other biodiversity data resources and delete potential duplicates from the final gazetteers. In the final project stage, we use the *CoL* webservice for entity linking.

2.3.2 Botanical Entity Recognition and Information Extraction

The recognition and extraction of botanical entities has been addressed with a strong focus on scientific Latin plant names. Several online tools have been made available to the research community to approach the difficult task of extracting domain-specific entities and associated information on different levels of granularity from text resources. In the following, we aim at giving a brief overview on existing

projects, methods, and resources in the field of botany and biodiversity informatics.

SPECIES [Pafilis et al., 2013]: The *SPECIES*-tool includes a dictionary-based approach to NER for the recognition of plant taxa in biomedical literature. Unlike rule-based systems using pattern-based sequential evidence (e.g. upper-case or lowercase, presence of typical Latin suffixes), dictionary-based tools mainly rely on string matching and dictionary-lookups. Especially when it comes to the recognition of multifaceted, language-specific vernacular names or alternative species names, dictionary-based tools outperform exclusively rule-based tools [Pafilis et al., 2013] (e.g. *TaxonGrab* [Koning et al., 2005] and *FAT* [Sautter et al., 2006]). They do, however, heavily rely on comprehensive plant name lists (also called gazetteers or dictionaries) of organism names that need to be constantly updated as soon as new scientific and vernacular entities are coined and used. As compared to other dictionary-based approaches, the *SPECIES* tool has the advantage of being much faster and more precise as compared to existing tools. Other dictionary-based approaches are *LINNAEUS* [Gerner et al., 2010], *AliBaba* [Plake et al., 2006], *Whatizit* [Rebholz-Schuhmann et al., 2008] and the *OrganismTagger* [Naderi et al., 2011]. As opposed to these dictionary-based systems, we only use the gazetteers to automatically create a silver-labeled training dataset. In addition, we sustain that botanical NER should not only be applicable to scientific text genres, but also to multilingual, heterogeneous text resources and genres.

Taxonfinder [Leary, 2014]: This online tool employs a dictionary-based approach to identify Latin taxonomic mentions of different taxonomic ranks such as phylum, class, order, family, genus and species. This feature is, in our eyes, especially useful when extracting hierarchical taxonomic relations and automatically constructing taxonomies. *Taxonfinder* is not only limited to the kingdom of plants, it also detects other scientific organism names. The detection of entities on a vernacular level has, however, not been addressed in this approach.

gnparser [Mozzherin et al., 2017]: This approach uses parsing methods to detect scientific plant names of varying degrees of structural complexity. In this manner, authorship information can be included as a valid component of more complex botanical entities. This tool can be valuable for taxonomists and biodiversity informaticians working with scientific plant names and information that can be derived from the attached authorship information. As opposed to the other approaches, the focus of this project mainly lies on correctly decomposing and analyzing the structure and composition of Latin plant names.

TNRS [Boyle et al., 2013]: The *Taxonomic Name Resolution Service* (TNRS) directly addresses the previously mentioned “names problem” [Boyle et al., 2013].

This approach uses name parsing and fuzzy matching to identify spelling errors, alternative spellings and outdated names and maps the variant to a currently accepted name version in a reference database. Primarily when dealing with taxonomic standardization and name normalization for international botanical standards, this approach represents a trustworthy tool to resolve synonyms, homonyms and spelling variants. While the *TNRS* can serve as a trustworthy resource for taxonomic standardization and name normalization, outdated spelling variants and synonyms on a vernacular level have not been considered.

GNA [Patterson et al., 2010]: The *Global Names Architecture* (GNA) is a valuable webservice to index scientific names and interconnect distributed information from different resources about species. The integrated global names resolver parses scientific input names and links them to a corresponding entry in a reference database or ontology. In our project, we use the more comprehensive Catalogue of Life (CoL) database and webservice to disambiguate and link the entity candidates.

NetiNeti [Akella et al., 2012]: Besides rule-based and dictionary-based approaches, different machine learning methods have been used to address the task of information extraction and text mining. The tool *NetiNeti* uses probabilistic machine learning methods, more specifically, Naïve Bayes and Maximum Entropy, to recognize scientific plant names and to discover new species names from text. The approach uses distributional evidence derived from contextual features and orthographic evidence to detect plant names in text. Similar to our approach, orthographic patterns and contextual evidence are used to predict whether or not a certain sequence of tokens corresponds to a plant name. Our neural models, conversely, do not rely on hand-crafted features and domain-specific context knowledge for feature extraction as in this approach. Consequently, the adaptation to new text genres and related domains might include work-intense feature adaptation and re-engineering. A promising aspect, that we also pursue in our work, is the recognition of name variants despite spelling and OCR-errors that are likely to occur especially when dealing with historical text material [Sharma et al., 2017].

Biomedical NER [Habibi et al., 2017]: Since neural models are capable of recognizing new, unseen entities that did not occur in the training data [Ni and Florian, 2016], they are particularly interesting for text mining and the extraction of new entities, for instance, new species names. The independence of neural models from hand-crafted, domain-specific features and their ability to generalize from examples and contextual information represents a considerable advantage over rule-based and dictionary-based approaches and feature-based machine learning methods [Lishuang Li et al., 2015; Ma and Hovy, 2016]. Habibi et al. [2017] apply the bi-LSTM-CRF ar-

chitecture proposed by Lample et al. [2016] to identify several biomedical entity types such as genes, proteins, diseases, chemicals and cell lines. In addition, they present an evaluation of the tagger's performance with regard to Latin species names that have been automatically annotated using the *SPECIES*-tool [Pafilis et al., 2013]. The examples presented in the analysis are, however, limited to non-botanical entities and, contrary to our approach, they do not apply fine-grained hierarchical entity labels. Another example for neural, domain-specific NER in the domain of bio-informatics has been presented by Lishuang Li et al. [2015]. This approach applies an extended recurrent neural network (RNN) to extract biomedical information which, in their case, is able to outperform CRF-based implementations. As opposed to our approach, no multilingual models are trained and the focus lies on scientific plant mentions on the taxonomic level of species.

2.4 Entity Linking

In the context of information extraction, the entity linking (EL) task has become increasingly important for several domains [Shen et al., 2015; Rao et al., 2013; McNamee and Dang, 2009; Bunescu and Pasca, 2006]. Entity linking or grounding, record linkage or entity resolution primarily involves mapping a sequence of tokens to a unique entry in a knowledge base for semantic disambiguation [Rao et al., 2013; Hachey et al., 2013; Cucerzan, 2007]. The detection and disambiguation of entity mentions in unstructured text is essential for the integration of knowledge derived from large text corpora and the subsequent population of knowledge bases [Dredze et al., 2010; Zheng et al., 2010]. Wikification approaches are based on Wikipedia as a backbone knowledge base in order to automatically link textual mentions to encyclopedic knowledge [Mihalcea and Csomai, 2007; Cucerzan, 2007]. Similar to the task of word sense disambiguation (WSD), EL deals with the resolution of lexical ambiguity of language [Moro et al., 2014; Navigli and Moro, 2014]. While WSD aims at matching a word form to an associated unique word sense in a reference dictionary, EL is concerned with linking textual mentions to entities or concepts in a reference encyclopedia [Moro et al., 2014; Navigli and Moro, 2014; Rao et al., 2013]. Due to anaphorical structures, synonyms, or abbreviated variants, the reference entity might not always exactly match the mention from the text [Moro et al., 2014; Rao et al., 2013]. As opposed to prevailing graph-based entity resolution approaches [Moro et al., 2014; Hachey et al., 2013, 2011], recently, neural end-to-end approaches have been proposed to jointly detect and link entity mentions in text [Kolitsas et al., 2018]. Oppositely, disambiguation-only approaches [Sheng et al., 2017] focus on the

disambiguation of gold standard named entities to a correct database entry. Raiman and Raiman [2018] propose *DeepType* as a multilingual system for entity linking by integrating symbolic knowledge. This knowledge comprises abstract type constraints, such as mutually exclusive types: IsHuman $\wedge$ IsPlant = {} (example from Raiman and Raiman [2018]). As a next step, they train a neural network on the formulated type system. Regarding domain-specific entity linking, approaches for gene, molecule, protein or organism normalization deal with the challenge of linking multifaceted entities to unique and internationally accepted identifiers in domain-internal databases to automatically extract structured knowledge from the current literature [Sheng et al., 2017; Morgan et al., 2008]. In summary, the core challenges of the EL task are name variations including abbreviations and alternative forms, mentions that match multiple identifiers in a reference database (entity ambiguity) and the absence of entity entries even in large knowledge bases [Shen et al., 2015; Rao et al., 2013; Zhang et al., 2010].

In the context of linked open data (LOD), aggregating and interlinking botanical and biodiversity knowledge is a powerful way to collaborate sustainably and to gradually populate international knowledge bases with existing and new, distributed resources [Minami et al., 2013]. The common resource description framework's (RDF) triple data structure guarantees machine-understandability, interoperability, and precise searchability [Bizer et al., 2008; Lehmann et al., 2015; Chiarcos et al., 2011; Minami et al., 2013]. Accordingly, data, results, and resources across multiple fields of research can be searched, extended, and re-used. Botany and biodiversity research commonly employ the simple knowledge organization system (SKOS) vocabulary [Miles et al., 2004], an RDF application, to model taxonomic concepts, semantic relationships, and properties of organisms. With this in mind, our project aims to enhance the automatic extraction and inter-linkage of vernacular and scientific entities. Further steps could include the automatic extraction of RDF triples from historical and current botanical literature, e.g. `daisy hasScientificName Bellis_perennis`, `daisy hasAlternativeName bruisewort` or also `authorXY usesVernacularName daisy`.

In the final stages of our project, we apply entity linking to disambiguate and interlink the entity candidates proposed by the language-specific neural models. As previously mentioned, we use the internationally accepted *Catalogue of Life* (CoL) webservice for the task of entity linking. The resolution of vernacular names and the subsequent mapping to a domain-specific database is, in our eyes, crucial to automatically access botanical knowledge and to enrich currently existing knowledge bases with naming variants and multilingual vernacular names.

3 Tagging Plants: Methods and Tools

In this chapter, we illustrate our overall approach and its substeps (see Figure 2). Regarding data collection, we describe the aggregation of extensive gazetteers (Section 3.1.1), which were used to automatically create the silver-annotations, and the digitization of historical botanical works in order to integrate different spelling variants and older vernacular names (Section 3.1.2). We then specify the language-specific text material used for the creation of four distinct datasets per language (Section 3.1.3). Subsequently, we present the tools and data formats used during data preparation and the individual steps involved in the preprocessing pipeline (Section 3.2). We then introduce our tailored dictionary-based annotation system and the pattern-based, semi-automatic corrections required for the final version of the annotated datasets (Section 3.3). In this context, we point out the importance of a manually corrected gold standard, necessary to assess and evaluate the true model performance (Section 3.4). We briefly sketch the preparatory steps involved in the successful application of the bi-LSTM-CRF architecture by Lample et al. [2016] that we will be using to train multiple models.

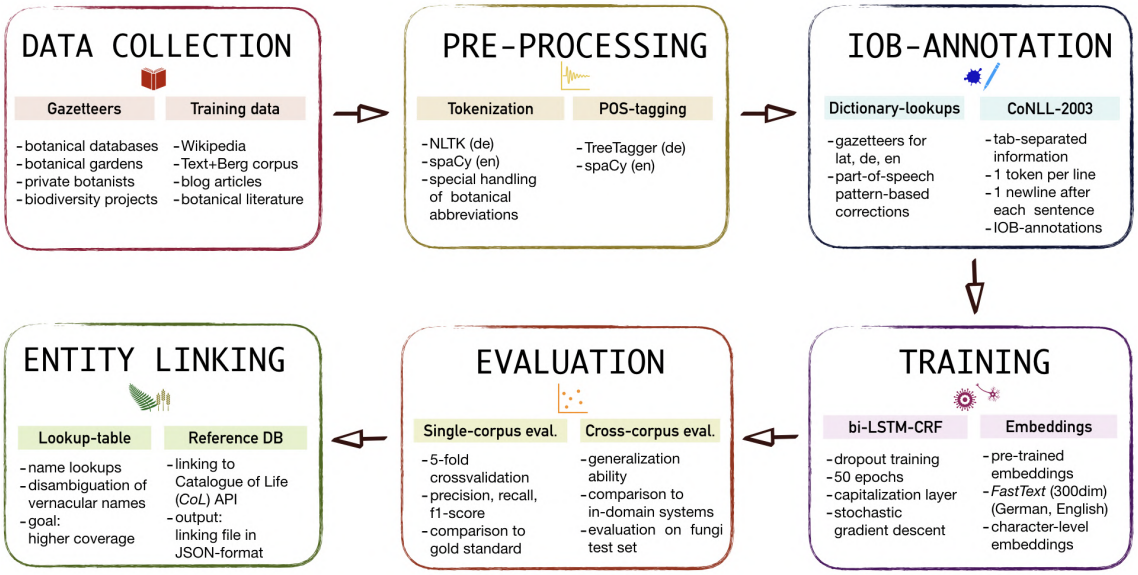

Figure 2: Methods, tools and resources used at different stages in this project.

3.1 Data Collection

3.1.1 Gazetteers

The manual annotation of training material is time-consuming and requires expert knowledge. Therefore, we compiled language-specific plant name lists, also referred to as *gazetteers*, *lexicons* or *dictionaries* - terms that are often used interchangeably in the context of named entity recognition (NER) [Nadeau and Sekine, 2007]. Since vernacular plant names are mainly concentrated on the taxonomic levels of species and family names, we divided the gazetteers for English and German into two distinct files for the entity classes *species* and *family* respectively.[1] The scientific plant names follow more systematic suffix patterns directly related to the taxonomic level they belong to and are thus easier to classify: The suffix pattern *-phyta* is typical for plant divisions (phylum) whereas the suffix *-opsida* is characteristic for plant classes. We divided the Latin names according to their specific suffixes into seven distinct hierarchical classes in separate files (`lat_species`, `lat_family`, `lat_genus`, `lat_subfamily`, `lat_class`, `lat_order`, `lat_phylum`). We removed all duplicate names from the final gazetteers, both within and across the single gazetteer files. In the case of an ambiguous duplicate item, we focused on its etymology (Latin or German/English) and on the potential existence of a semantically equivalent vernacular name. For instance, the Latin genus names *Adonis, Curcuma, Monarda, Rhododendron, Stevia* are frequently used as such in German. Their etymological origin is the Latin form and equivalent vernacular names exist in German as well: *Adonisröschen* ('yellow pheasant's eye'), *Kurkuma*, *Kurkume* or *Gelbwurzel* ('turmeric'), *Monarde* ('crimson beebalm'), *Rhododendren* ('rhododendron'), *Süßkraut*, *Süßblatt*, or *Honigkraut* ('sweetleaf').

In total, we created seven gazetteers for Latin plant names and two separate files each for German and English. The unique gazetteer file name serves as an annotation class label and is combined with the corresponding positional information (I for inside, O for outside and B for beginning [see Section 3.3]). An example: The system detects the token *Amygdalus* in a sentence, which matches an entry of the Latin genus gazetteer. Accordingly, it assigns the tag `B-lat_genus`.

Table 2 illustrates the distribution of the plant names including the generated variants per gazetteer and per language. It should be mentioned that the number of German family names is higher due to the automatic insertion of possible variants

[1] Vernacular names on the taxonomic level of subfamilies or phylum are rarely verbalized in a language. On the taxonomic level *order*, one can find German expressions such as "Zaubernußartige" (Hamamelidales). We included such cases in the vernacular family gazetteer.

(see Subsection 3.1.1.3). Furthermore, the English language frequently employs Latin family names instead of vernacular plant names, which explains the smaller number of family names as compared to the German and Latin gazetteers.

	species	family	genus	subfamily	class	order	phylum
Latin	2,152,656	1,292	31,617	262	47	203	25
German	63,932	3,873	/	/	/	/	/
English	67,124	444	/	/	/	/	/

Table 2: Gazetteer sizes for vernacular (*species* and *family*) and scientific names (*species*, *family*, *genus*, *subfamily*, *class*, *order* and *phylum*).

For the sake of clarity, we would like to stress that the large number of Latin species also includes abbreviated forms such as *B. perennis* (Bellis perennis).

3.1.1.1 Scientific Gazetteers

For the creation of the scientific gazetteers, we aggregated multiple large freely accessible and up-to-date check lists or downloadable archives provided by institutions such as the *Catalogue of Life* (CoL) [Roskov et al., 2018], the *Global Biodiversity Information Facility* (GBIF) [GBIF, 2018], the *International Plant Name Index* (IPNI) [IPNI, 2012], and the *Multilingual Multiscript Plant Name Database* (MMPND) [Porcher, 1999]. The *CoL*, for example, offers the *Darwin Core Archive format* to export parts of the database. We parsed the resulting data structure and automatically stored each name on a single line.[2] These comprehensive, albeit not complete, botanical databases do not only list currently accepted Latin names, but also include outdated synonyms or accepted spelling variants. For better coverage, we included all possible variants in the final Latin gazetteers, since discarded synonyms might occur especially in non-scientific text genres such as blog articles or historical literature.

3.1.1.2 Vernacular Gazetteers

The creation of the German gazetteer has proven to be more challenging: Due to the lack of large, comprehensive gazetteers including synonyms and spelling variants in German-speaking regions, we combined several structured (tab-separated or

[2]See Appendix C for the Python scripts used to create the gazetteer files.

comma-separated files), semi-structured (Wikipedia articles) and un-structured data resources (digitized botanical works). Besides multiple small resources kindly provided by private experts and botanists, we also retrieved numerous common names from the trivial names section (if existent) in the Wikipedia category "Vascular plants of Germany" using the API.[3] Other valuable resources were provided by several institutions in German-speaking regions concerned with biodiversity such as *Info Flora* [Aeschimann and Heitz, 2005], local botanical gardens or private botanists. For English, we used the *Multilingual Multiscript Plant Name Database* (MMPND) [Porcher, 1999] to retrieve additional vernacular names. Subsequently, we used the *Catalogue of Life* (CoL) [Roskov et al., 2018] to extract the available vernacular names for German and English. We noticed, however, that this resource does not employ language tags consistently. For instance, *grassflower* has been tagged as [English] and *pond cypress* as [Eng]. Moreover, German language tags were often missing in the *CoL* archives. For this reason, we used the `langid`-module for Python [Lui and Baldwin, 2012] to double-check the original language of the vernacular name. Other valuable resources for this project stage were the vernacular names provided by the botanical databases *Germplasm Resources Information Network* (GRIN) [Wiersema, 2018] and *Global Biodiversity Information Facility* (GBIF) [GBIF, 2018] as well as the botanical dictionary entries kindly provided by the authors from the online dictionary *dict.cc*. Lastly, we also extracted dialectal and partially archaic variants of common plant names. For this purpose, we processed and, if necessary, digitized several botanical works.[4] In general, we avoided works in blackletter typeface and manual post-corrections of the retrieved plant names but especially when dealing with historical data and potential OCR-errors, some manual corrections to ensure high-quality gazetteers were inevitable. Finally, we processed, cleaned, and stored each resource separately in order to exclude noisy resources at a later stage. Thus, we avoided the potential infiltration of noisy plant names into the final gazetteers.

3.1.1.3 Generation of Name Variants

To increase coverage, we automatically added possible name variants to the gazetteers. For the Latin entries, we generated the associated, abbreviated variants, which are likely to occur in scientific text genres (see Example 1). For German, we created additional variants by attaching or removing morphological endings and by splitting

[3] Original title: "Liste der Gefässpflanzen Deutschlands" [Wikipedia, 2018c]. For English, we used the Wikipedia page "List of plants by common name" to retrieve additional vernacular names.

[4] See Appendix A for a full list of botanical works used for this project.

compounds into their single components (3). To avoid the insertion of potentially noisy data, we only covered the most frequent and systematic variants. This includes adding morphological suffixes to German plant family names that usually end with the form -*Gewächse* ('plants') or -*Familie* ('family') (2). Additionally, we merged hyphenated compounds (3) resulting from the synthetic nature of the German language and split the descriptive first component.[5] The same applies to plant names containing a descriptive adjective in the first position (4).

	Plant name	**Translation**	**Generated variants**
1.	*Eugenia floccosa*	a species of myrtle	E. floccosa
2.	*Sauergrasgewächse*	sedges	Sauergras-Gewächse, Sauergras-Gewächses, Sauergras-Gewächsen, Sauergras-Gewächs, Sauergrasgewächse, Sauergrasgewächses, Sauergrasgewächsen, Sauergrasgewächs
3.	*Vogel-Sternmiere*	chickweed	Vogelsternmiere (merged) Sternmiere (split)
4.	*Johannisbeere, Schwarze*	blackcurrant	Schwarze Johannisbeere (inverted full species name) Johannisbeere (only genus name)

Table 3: Automatically generated name variants (abbreviations, split or merged compounds, morphological variants) for German and Latin gazetteers.

3.1.2 Digitization of Botanical Works

As described in the previous section, precompiled, machine-readable gazetteers for German plant names and their synonyms are not easy to find, particularly regarding outdated spelling variants and historical synonyms. Therefore, we digitized multiple historical botanical works in order to extract the vernacular names and variants used by the author (see Appendix A). Other digitized works have been kindly provided by *plazi* (Bern, Switzerland). We used the ABBYY ® FineReader software to extract the text from the scanned works in PDF format. The PDFlib TET 5.1 software package was a valuable tool to extract the XML or the raw text data from already OCRed PDF documents. Depending on the quality and structure of the resulting digitized version, we applied further customized processing steps to clean, extract, and store the present plant names. Some works provide schematic name listings

[5] Pahler and Rücker [2001] published a set of rules to standardize the spelling of German plant names by splitting the name part designating the genus with a hyphen. The writing style in historical literature or by non-expert botanists, however, proves that the existing spelling variants are often innumerable.

that can be parsed using simple pattern matching (see Figure 3). In other cases, we simply used the name index to extract the plant names.

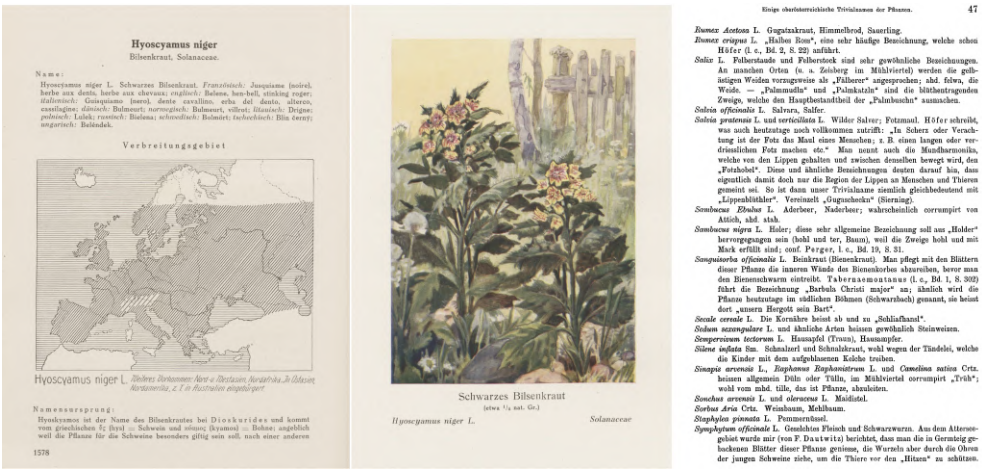

Figure 3: Example for digitized botanical works from Madaus [1938] (left) and the name index from Pfeiffer [1898] (right).

3.1.3 Training Corpora

In order to explore the behavior of neural named entity recognition with regard to cross-corpus and genre-adaption performance, we collected training material from four different text genres and writing styles. While scientific botanical literature almost exclusively relies on the use of explicit Latin plant names or equivalent abbreviations, other text genres such as blog articles or mountaineering reports employ a more informal and impulsive writing style and thus tend to use vernacular plant names more frequently. We applied a modular design and processed each data resource independently to train separate models for each data resource and text genre. This allows us to explore the impact of different parameter combinations on each dataset and to analyze the cross-corpus generalization performance of the neural models. For comparison, we created a supplementary combined dataset from all the available resources. Table 4 reports the overall details for each corpus including the number of tokens, types, and sentences.

Wiki Corpus: We created a corpus from German and English Wikipedia abstracts describing plants of the vascular plant category. For this purpose, we iterated over each plant listed in this Wikipedia category and used the API and a Python wrapper provided by Majlis [2017] to retrieve the introductory abstract for each article. Scientific names are usually combined with vernacular names within one sentence, which makes this resource particularly interesting and challenging for our project.

	Wiki		TB		PlantBlog		BotLit/S800		Total/Average	
domain	*Wikipedia articles*		*mountaineering reports*		*blog articles*		*botanical literature*			
	German	English	German	English	German	English	German	English	German	English
number of tokens	330,119	330,495	52,265	5,359	13,212	17,289	886	24,773	396,428	377,916
number of types	34,384	27,354	13,319	2,022	3,971	3,392	493	4,203	52,167	36,971
mean token length	5.42	4.42	5.06	4.45	5.47	4.17	5.25	5.07	5.05	4.52
number of sentences	13,882	15,280	2,289	153	720	876	42	960	16,933	17,269
mean sentence length	23.78	21.62	22.83	35.02	18.35	19.73	21.09	25.80	21.51	25.54

Table 4: Corpus details for German and English datasets.

In this text genre, a systematic and structured writing style is usually employed when describing a plant species or other taxonomic levels (see Table 5). Frequently, listings of vernacular names are preceded by an explicit expression following the patterns "common names include [...]", "also known by the common name [...]'" or "the vernacular name is [...]".[6]

PLANT NAME	**is_a**	**DESCRIPTION**	**of_the**	**FAM_NAME (or higher taxa)**
Bellis perennis	is a	common European species of daisy	of the	*Asteraceae family.*
Die Gänseblümchen (Bellis)	sind eine	Pflanzengattung	aus der	*Familie der Korbblütler (Asteraceae).*

Table 5: Characteristic Wikipedia writing style when describing botanical entities.

TB Corpus: The Text + Berg corpus is an annotated linguistic corpus consisting of digitized yearbooks of the Swiss Alpine Club (SAC) with mountaineering reports in German/French/Italian/Romansh (SAC yearbooks), English (Alpine Journal), and French (Echo des Alpes) [Volk et al., 2010]. For this corpus, we applied our previously created gazetteers to select a subset of sentences from the German and English part of the corpus release 151 [Bubenhofer et al., 2015]. This sub-selection ensures that each sentence contains at least one scientific or vernacular plant name mention and hence guarantees a high density of botanical entities in this training set. For German, we extracted sentences from the time span of 1970 to 2015, whereas for English, we only used material from 1980 to 2008 due to the restricted availability of the recent years and the lower OCR quality of the older yearbooks. The writing style is generally informal and, depending on the botanical expertise of the author and the publication period, different types of plant names are used (see Table 6). An author

[6]The same applies to the German Wikipedia dataset where patterns, such as "gebräuchliche Volksnamen sind [...]" ('common trivial names are'), "nur regional gebräuchlich sind die Trivialnamen [...]" ('the following trivial names are only locally known [...]') or "weitere Trivialnamen sind [...]" ('additional trivial names are [...]') occur frequently.

might use, for instance, both vernacular and scientific names within one sentence (1, 5), exclusively vernacular names (2, 6) or exclusively scientific names (rather for Alpine Journal or older German yearbooks) (4, 7). Frequently, the writers combine vernacular and scientific names in extensive listings (3, 4).

	Sentence/ ***Translation***	**Source**	**Year**
1.	Um so schöner sind [...] die Wiesen mit Alatau-Schwingel (Festuca alatavica) und der Segge Carex tristis. '*The meadows with Festuca alatavica and the Carex tristis sedge are even more beautiful.*'	Swiss Alpine Club (SAC) Text+Berg corpus	1970
2.	Man [...] erblickt einen hellen, schlanken Campanile, daneben die schwarze Flamme einer Zypresse. '*One sees a bright, slim campanile, next to the black flame of a cypress.*'	Swiss Alpine Club (SAC) Text+Berg corpus	1971
3.	Bors Segge (Carex Born), Altai-Hungerblume (Draba altaica), Himalaya-Edelweiss (Leontopodium leotopodinum), das zierliche Gras Colpodiuum himalaicum und das Seidenhaarige Fingerkraut (Potentilla serìcea). '*Born's sedge (Carex Born), nailwort (Draba altaica), Hamalayan edelweiss* *(Leontopodium leotopodinum), the delicate weed Colpodiuum himalaicum* *and the silk cinquefoil (Potentilla serìcea).*'	Swiss Alpine Club (SAC) Text+Berg corpus	1970
4.	These include several species of Rhododendron, Hypericum, Buddleja, Syringa, Berberis, Daphne and the delightful Clematis montana, and much more besides.	Alpine Journal Text+Berg corpus	1991
5.	A widespread starry white-flowered mouse-ear (Cerastium trigynum), a rather elegant gentian (G. stricta) [...].	Alpine Journal Text+Berg corpus	1983
6.	The trees were the real enemy, old trees, young trees, rotten and burnt trees, Douglas Firs, Jack Pines, Spruce, Poplar and Birch.	Alpine Journal Text+Berg corpus	1990
7.	The island, called Islota Caro, was covered in Desfontainia spinosa [...].	Alpine Journal Text+Berg corpus	2008

Table 6: Example sentences for plant-related mountaineering reports from the Swiss Alpine Club and the Alpine Journal (*TB* corpus [Bubenhofer et al., 2015]).

For our purposes, this corpus is highly interesting as it combines a descriptive and poetic language with a passion for botany. Due to the diachronic dimension of the corpus, the writing style is highly individual and depends on the specific author and the publishing year. The Latin names used by the authors do not, however, always coincide with the currently accepted names and are often outdated synonyms.

PlantBlog Corpus: This corpus consists of blog articles retrieved from the Internet about plant-related topics such as gardening, phytotherapy, herbalism, or household remedies. Blog authors usually employ an informal writing style and tend to use vernacular instead of scientific plant names since not all potential readers might be botanists or botanically-minded. An example of typical blog-writing style for German (see Example 1) and English (2) regarding ginger ('Zingiber officinale'):

1. Vom Ingwer wird die Wurzel verwendet, er enthält überdurchschnittlich viele ätherische Öle.
   'You can use the ginger's root, it contains an above-average amount of essential oils.'

2. The pain-relieving potential of ginger appears to be far-reaching.

BotLit/S800 Corpus: For the German part of this data resource, we manually selected text passages from historical botanical literature containing a high concentration of either scientific or vernacular plant names. Although being less extensive, this dataset allows testing the performance of our models regarding old spelling variants or archaic names. We included text excerpts from Spescha [2009] containing regional descriptions of flora and fauna, originally published in 1806. We also used text snippets from Höhn-Ochsner [1986] and Bosshard [1978] containing folk sayings with old vernacular names in several local dialects. The German part of the *BotLit* corpus is particularly interesting because of the frequent use of non-standard names, dialectal variants, and old spellings. For instance, Spescha [2009] uses "Weisthanne" instead of the currently accepted forms "Weißtanne" or "Weiß-Tanne"for the *European silver fir* (Abies alba).

1. Unter dem Nadelholz zeichnet sich die Roththanne aus, [...] der Weisthannen sind wenig [...]. [Spescha, 2009, 144-145]
   'Among the conifers European spruces are frequent, but there are only a few silver firs.'

2. Rübli säe im Fisch und nidsichgehenden Mond [...]. [Höhn-Ochsner, 1986, 63]
   'You should sow the carrots in the zodiac sign pisces and when the moon is waning.'

For English, on the other hand, we used the *S800* corpus created by Pafilis et al. [2013]. This data resource contains abstracts from different scientific fields such as medicine, virology, and botany. For our experiment, we only used the subsection based on botanical abstracts, which almost exclusively contains Latin plant name mentions. While vernacular names are less frequent in this corpus, scientific names occur in diverse shapes, including abbreviations (1), extensive subspecies' or varieties' names (2).

1. Seventeen natural populations of A. thaliana were geo-referenced in north-eastern Spain [...].

2. Selenium (Se)-fortified broccoli (Brassica oleracea var. italica) has been proposed as a functional food for cancer prevention [...].

Table 7 illustrates the distribution of the annotated absolute and unique entity class labels per language and dataset. For both languages, the usage of Latin plant names from higher taxonomic levels such as class, order, and phylum is restricted to the Wikipedia dataset. Text genres including mountaineering reports (TB corpus) and blog articles tend to use vernacular names on the level of species and family as well as Latin species, genus, and family names.

	Wiki				**TB**				**PlantBlog**				**Botlit**			
	German		English		German		English		German		English		German		English	
	all	unique	all	unique	all	unique	all	unique	all	unique	all	unique	all	unique	all	unique
de/en_species	21,957	7,597	11,674	3,269	2,954	767	275	188	743	425	740	317	67	64	171	44
de/en_fam	6,314	358	1,897	152	123	29	4	3	25	14	12	4	0	0	0	0
lat_species	15,722	4,782	8,244	3,820	467	286	301	227	133	106	107	69	32	30	267	100
lat_genus	4,957	1,422	4,569	2,717	105	66	74	50	35	29	45	28	4	4	54	12
lat_fam	5,453	278	2,335	422	26	20	1	1	2	2	3	2	0	0	5	3
lat_subfam	565	70	255	125	0	0	0	0	0	0	0	0	0	0	1	1
lat_class	110	8	32	17	0	0	0	0	0	0	0	0	0	0	0	0
lat_order	281	60	183	78	0	0	0	0	0	0	0	0	0	0	0	0
lat_phylum	11	4	28	13	0	0	0	0	0	0	0	0	0	0	0	0
total	**55,370**	**14,579**	**29,217**	**10,613**	**3,675**	**1,168**	**655**	**469**	**983**	**576**	**907**	**420**	**103**	**98**	**497**	**159**
		26%		36%		31%		71%		61%		46%		95%		31%

Table 7: Fine-grained classes of botanical entities annotated in each corpus.

3.2 Linguistic Preprocessing

In order to format and enrich the datasets with linguistic information before annotating the botanical entities, we applied several preprocessing steps. First, we segmented the raw text data into single sentences and performed tokenization using the *Natural Language Toolkit* (NLTK) [Loper and Bird, 2002] for German. We then applied the *TreeTagger* [Schmid, 1995] to associate a lemma to all known tokens and a <unknown> tag to all unknown tags. Simultaneously, we stored the respective part-of-speech (POS) tag on the same line. For the linguistic annotation of the English datasets, we used the tokenizer and part-of-speech tagger of *spaCy* [Honnibal and Montani, 2017], an NLP library for Python. For comparison, we also tested the *spaCy* POS-tagging performance on the German data, but the tagging behavior on German vernacular plant names has proven to be unsystematic.

3.2.1 Treatment of Botanical Abbreviations

Current NLP tools such as *NLTK* or *spaCy* are prone to errors when processing domain-specific text genres, e.g. botanical literature. Typical botanical abbrevia-

tions such as var., convar., ssp., subsp., and others[7] are frequently treated as sentence boundaries and are therefore split. Since these abbreviations are often part of Latin plant names, e.g. in *Cannabis sativa var. spontanea* ('hemp') or *Daucus carota subsp. carota* ('wild carot'), this circumstance results in erroneous sentence segmentation and truncated plant names. We corrected this unexpected behavior using a regular expression to merge the plant names and the abbreviations (see Section 4.6.1).

3.2.2 The CoNLL-2003 format

For the subsequent training of the LSTM-CRF models, we used the CoNLL-2003 format [Tjong Kim Sang and De Meulder, 2003] to store the training corpora. This data format is commonly used to approach the task of named entity recognition and allows the association of multiple tags per data row and per token. The linguistic information resulting from the preprocessing steps described above is stored in one row (token, lemma, pos-tag) and separated by tabs while newlines represent sentence boundaries (see Table 8). The rightmost entity tag encodes the annotated entity label in IOB format [Tjong Kim Sang and De Meulder, 2003] (see Section 3.3.1).

3.3 Dictionary-based Annotation

The task of accurately identifying taxa in natural language has been approached using rule-based systems exploiting the schematic Linnaean name patterns or dictionary-based systems using name lookups and string matching [Pafilis et al., 2013]. To expand such current approaches from the recognition of merely Latin names to vernacular names and alternative spellings, we integrated a tailored dictionary-based annotation system into the annotation process. The shape and structural complexity of vernacular plant names does not follow any systematic rules, e.g. *Unechter Veränderlicher Gold-Hahnenfuß* (Ranunculus pseudovertumnalis). In this example, the first two tokens are adjectives that might also occur in other, non-botanic contexts. Except for the characteristic uppercase, which is typical for all German nouns, no systematic pattern matching rules can be applied to extract vernacular names from text. Equivalently, multiword expressions are frequently used in English to refer to plants, e.g. *hen and chicks* or *hen-and-chickens* (Jovibarba globifera and other species).

[7] We considered the following botanical abbreviations (*var., convar., agg., ssp., sp., subsp., x., L., auct., comb., illeg., cv., emend., al., f., hort., nm., nom., ambig., cons., dub., superfl., inval., nov., nud., rej., nec., nothosubsp., p., hyb., syn., synon.* (after Meades [2018]).

First, the dictionary-based annotation system checks if any name from the gazetteers occurs within the current sentence. Next, it assigns the language tag (de, en or lat) together with the correct taxonomic label (species, family, genus, subfamily, class, order, or phylum) depending on the gazetteer file where the matching token (or sequence of tokens) has been found. To store the information of the plant entities found in a sentence, we used the IOB (Inside, Outside, Beginning) scheme for annotation (see Table 8).

TOKEN	LEMMA	POS-TAG	IOB-TAG
Der	die	ART	O
Gefleckte	gefleckt	ADJA	**B-de_species**
Schierling	<unknown>	NN	**I-de_species**
Conium	<unknown>	NN	**B-lat_species**
maculatum	<unknown>	NN	**I-lat_species**
gehört	gehören	VVFIN	O
mit	mit	APPR	O
dem	die	ART	O
Wasserschierling	<unknown>	NN	**B-de_species**
zu	zu	APPR	O
den	die	ART	O
Doldenblütlern	<unknown>	NN	**B-de_fam**

Table 8: Dictionary-based annotation of the datasets with language-specific labels and taxonomic information encoded in the IOB tag.

3.3.1 The IOB tag scheme

The IOB tag scheme is widely used to represent single tokens or a sequence of tokens that can be grouped into non-overlapping and non-recursive text chunks, thus constituting a semantic unit [Ramshaw and Marcus, 1995; Sang and Veenstra, 1999; Tjong Kim Sang and De Meulder, 2003]. This annotation scheme indicates the beginning (B), inside (I) and outside (O) of a named entity identified within a sentence. Since Lample et al. [2016] did not observe significant improvements when using the extended IOBES scheme, which also marks singleton entities (S) and endings (E), we considered the IOB scheme to be sufficient for our experiment. We made sure that the longest possible sequence constituting a plant name is annotated instead of possible consecutive subsequences. In the following example, our annotation system identifies "Datura stramonium" as the longest possible sequence and assigns two consecutive tags, namely `B-lat_species` followed by `I-lat_species` (Example 1),

instead of the possible subsequence "Datura" found in the `lat_genus` list. The same applies to the annotation of vernacular names: The German species name "Fiederschnittige Perowskie" represents a semantic unit even though it would be possible to encounter the singleton "Perowskie" in other contexts (2).

1. ***Datura stramonium*** *ist eher selten zu finden.*
   '**Datura stramonium** can be rarely found.'

2. *Die* ***Fiederschnittige Perowskie*** *ist ein Halbstrauch mit lilablauen Blüten.*
   '**Perovskia abrotanoides** is a dwarf shrub with lilac-blue flowers.'

3.3.2 Pattern-Based Corrections

To ensure a higher quality of the training material, we semi-automatically corrected vernacular multiword plant names that were systematically missed by our annotation system due to morphological endings, structural complexity, or incomplete gazetteers.[8] For this purpose, we applied a regex search based on the part-of-speech annotations to detect uppercase adjectives that were tagged as O (outside) and precede a noun tagged as vernacular plant name (`B-de_species`):

regular expression	\n([A-Z]\w+\t.*?\tADJA\t)O\n(.*?\t)B-de_species\n
replace pattern	\n\1B-de_species\n\2I-de_species\n

Table 9: Find and replace pattern to systematically detect and annotate multiword plant names missed by the dictionary-based annotation system.

For English, hyphenated compound names were a frequent reason for erroneous or partial annotations, e.g. "yellow star-of-Cyprus" (*Gagea juliae*), "five-seeded plume-poppy" (*Macleaya cordata*), "small-flowered touch-me-not" (*Impatiens parviflora*) or "African weed-orchid" (*Disa bracteata*). In the final version of the annotated datasets, we corrected these misannotations using either regular expressions, part-of-speech patterns or manual corrections to guarantee high-quality training data.

[8] While the gazetteers comprise most generic nouns such as "Hahnenfuß" (Ranunculus), the species members might not always be complete especially regarding rare species such as the "Wolliger Hahnenfuß" or 'downy buttercup' (Ranunculus lanuginosus).

3.4 Creation of Gold Standard

To evaluate the performance of the neural models on a high-quality gold standard, we manually corrected one test fold of the silver-labeled combined dataset. The combined dataset comprises random sentences from all four corpora. The size of one test fold, in this case, equals to 76,783 tokens for German and 75,156 tokens for English and represents approximately 20% of the combined dataset. Manual annotation is highly time-consuming and requires expert knowledge in order to decide on context-dependent and ambiguous cases (Example 1). For instance, in some cases "Rhododendron" could be tagged as `B-de_species` (2) while in other contextual circumstances, the name should be tagged as `B-lat_genus` (3):

1. Das ist eine Pflanzenart aus der Gattung der **Sonnenwenden** (Heliotropium).
   'This is a plant species from the genus Heliotropium.'

2. Ein kleiner **Rhododendron** macht sich hier ausgezeichnet.
   'A tiny rhododendron fits in perfectly.'

3. Alpenazaleen (**Rhododendron**) kommen oberhalb der Baumgrenze vor.
   'Snow-roses grow above the tree line.'

In total, we devoted approximately 10 hours to the manual annotation of each language-specific gold standard. During evaluation, we applied the neural model trained on the first training fold of the combined dataset in order to ensure that none of the gold test sentences have already been seen during training. We present a detailed comparison between the automatically annotated silver-standard and the manually corrected gold standard in Section 4.1. This juxtaposition of the model performance on silver-labeled and gold-labeled data allows us to additionally assess the performance of the dictionary-based annotation system at different stages.

3.4.1 Annotation Guidelines

We applied general annotation rules to ensure consistent annotations throughout the gold standard for German and English. First, we decided not to annotate German compounds such as "Anis-Geruch" ('anise smell') or "Tomatenfrüchte" ('tomato fruits'), as the entire token does not refer to a plant in this case. Furthermore, we consistently annotated the longest possible variant of a complex plant name as in *Lactuca sativa var. crispa* ('curled lettuce'). In other cases, we trusted our linguistic expertise to disambiguate difficult cases depending on the context. As demonstrated in Example 1, the German expression "Sonnenwende" ('solstice') also constitutes a

vernacular name to denote the *European turnsole* (Heliotropium europaeum).[9] If necessary, we double-checked the existence of peculiar dialectal variants in German such as "Stoh up und gah weg" used for the *European centaury* (Centaurium erythraea) or "Traut Babbichen sieh mich an" for the *common moonwort* (Botrychium lunaria). Similarly, we verified the existence of exotic English vernacular names such as "Roan Mountain false goat's beard" (Astilbe crenatiloba) or "Jack go to bed at noon" (Tragopogon pratensis) in botanical reference works and annotated the longest possible sequence of tokens in the gold standard datasets.

3.5 Application of the bi-LSTM-CRF Architecture

For our experiments, we applied the bidirectional LSTM-CRF architecture proposed by Lample et al. [2016] to train multiple language-specific neural models for German and English.[10] To this end, we trained single-dataset models on the automatically annotated datasets (*Wiki*, *TB*, *PlantBlog*, *BotLit/S800*) and a combined-dataset model for both languages. To avoid overtraining, we applied 5-fold cross-validation (80% training set, 20% test set) and used the average scores over all folds for final model comparison (see Section 4.2). The corpora were split into individual folds using the Python package *scikit-learn* [Pedregosa et al., 2011]. Due to the small size of datasets such as the *PlantBlog* corpus or the German *BotLit* corpus, we did not use a supplementary development set during training. In Section 4.2, we give a detailed overview of the different parameter combinations used for training such as number of epochs, word embedding dimension, integration of pre-trained embeddings, hidden layer size, and character embedding dimension.

[9] Examples of sequences of tokens that formally refer to plants, but depending on the context they may have a different meaning: "Baumwolle" ('cotton': plant vs. fabric), "Bergröte" ('dyer's woodruff': plant vs. afterglow).

[10] All neural models have been trained on the GPU-server of the Institute for Computational Linguistics, Zurich.

4 Neural Models: Results and Evaluation

In this chapter, we present the results of our semi-supervised approach towards neural named entity recognition (NER) for botanical and biodiversity contexts. First, we discuss the overall quality of the automatic annotations by comparing the manually corrected gold standard and the corresponding silver standard annotations (see Section 4.1). We then present an overview of the parameters and dataset combinations used for training before moving on to a detailed model evaluation from a single-dataset and a cross-dataset perspective (Sections 4.2 and 4.3). We discuss persisting sources of errors in the light of two subsequent training runs and explore manifold factors such as the adoption of specific parameters, the quality of the training data, language-specific peculiarities, and the relative performance on the gold standard data.

4.1 Evaluation of Semi-Automatic Annotations

In our experiments, we distinguish between two subsequent annotation and training rounds. The initial version of the four training corpora after the first annotation round, for instance, includes partial annotations or misannotations due to interfering factors: Erroneous sentence segmentation caused by botanical abbreviations, structurally complex named entities including hyphenated compounds, or multiword expressions (MWE) and language-specific ambiguity. During the second annotation round, we semi-automatically corrected these systematic sources of error based on part-of-speech (POS) patterns or by re-merging those sentences that have been split by the tokenizers using regular expressions. Last but not least, we double-checked and, if necessary, corrected the most frequent and eye-catching ambiguous cases. In the Sections 4.6.2 and 4.6.3, we conduct a detailed error analysis and discuss structurally complex and ambiguous cases for German and English.

An evaluation of the dictionary-based annotations led to the following results: For German, the initial comparison of the dictionary-based annotations against the manually corrected gold standard, showed an overall annotation accuracy of 95.44% with an F_1-score of 85.05% (see Table 10). In contrast, the final dictionary-based annotation system in combination with semi-automatic, pattern-based corrections improved the overall annotation performance with an accuracy of 98.03% (F_1-score 93.70%). For the English dataset, the dictionary-based system achieved an accuracy of 97.55% with an F_1-score of 84.88% in the first round. After the second annotation round, the evaluation showed an improved accuracy of 98.59% (F_1-score 91.80%).

	German				**English**			
	A	P	R	F	A	P	R	F
1st annotation round	95.44	89.10	81.36	85.05	97.55	90.95	79.57	84.88
2nd annotation round	98.03	96.84	90.76	93.70	98.59	94.58	89.19	91.80

Table 10: Accuracy (A), precision (P), recall (R) and F_1-score (F) for dictionary-based annotation system resulting from the direct comparison between silver-labeled data and manually corrected gold standard.

The remaining mistagged or entirely missed instances are, in general, caused by language-specific ambiguity or context-dependent usages. For the morphologically rich language German, another frequent source of error are morphological suffixes. As opposed to the German family name variants automatically added to the gazetteers during the data collection stage, the species gazetteer usually only comprises the base form of a species name. In the case of the toxic legume "Berg-Spitzkiel" ('locoweed', *Oxytropis montana*), the nominative base form is present in the gazetteer, but the plural "Berg-Spitzkiele" or the genitive form "Berg-Spitzkiels" are missing. Thus, the dictionary-based annotation system fails to correctly annotate such cases. To explore the generalization capacity of the neural models, we avoided the integration of hand-crafted rules and manual corrections to match morphological variants. Latin species names including additional information on the infraspecies level, thus spanning over multiple tokens, are another common source of error, e.g. *Carex scirpoidea subsp. convoluta* (example taken from Mozzherin et al. [2017]). To sum up, the integrated dictionary-based annotation system achieved a satisfying performance with F_1-scores of >80% in the first annotation round and >90% in the second annotation round, as evaluated on the manually corrected gold standard for both languages.

4.2 Individual Dataset Evaluation

In this section, we evaluate the model performance on the four corpora (*Wiki*, *TB*, *PlantBlog*, *BotLit/S800*) in a single-corpus setting using the bi-LSTM-CRF architecture proposed by Lample et al. [2016]. We will assess the in-genre performance for the single text types and explore the impact of manifold training parameters. After introducing and motivating the baseline system in Section 4.2.1, we evaluate the model performance across different text genres using 5-fold cross-validation. More specifically, we discuss the resulting impact on the performance when using pre-trained embeddings (`pre_emb`) (Section 4.2.2), dropout training for better generalization ability (`dropout`) (Section 4.2.3), an augmented character embedding dimension (`char_dim`) (Section 4.2.4) and a capitalization feature dimension (`capdim`) (Section 4.2.5). We train a single-dataset model for each training dataset (*Wiki*, *TB*, *Blogs*, *BotLit/S800*) and a mixed-dataset model trained on the combined resources.

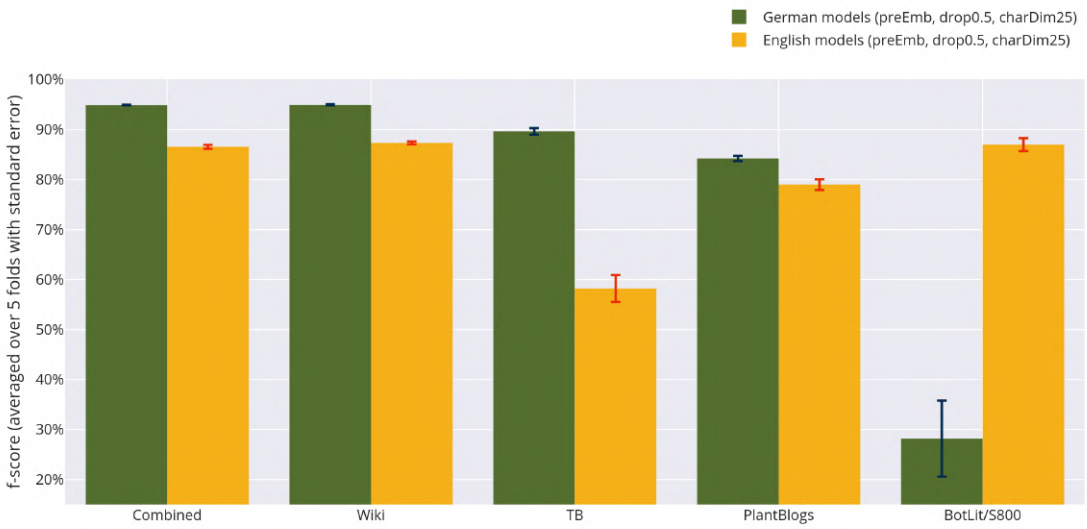

Figure 4: F_1-scores per dataset for single and combined models (pre-trained embeddings, dropout 0.5, char-embedding dimension 25). The error bars represent the standard error computed over the 5-fold cross-validation scores.

Figure 4 displays the average F_1-scores for the individual and the combined datasets based on the models trained using the 300-dimensional pre-trained *FastText* word embeddings [Grave et al., 2018], a balanced dropout rate of 0.5 and a character embedding dimension of 25. The error bars represent the standard error for each average F_1-score over the five cross-validation folds. As visible in Table 11, the observable performance variability over the five folds is small for larger datasets

such as the *Wiki* dataset or the combined dataset. Oppositely, we observed a large variability of ± 7.61 for the German *BotLit* dataset (M = 28.19, SD = 17.01), which contains only 42 sentences with partially historical name variants.

	German				English		
	F_1-score (AVG)	SD	SE		F_1-score (AVG)	SD	SE
Combined	94.92	0.12	0.05	**Combined**	86.55	0.87	0.39
Wiki	94.96	0.24	0.11	**Wiki**	87.35	0.62	0.28
TB	89.65	1.45	0.65	**TB**	58.20	6.03	2.70
PlantBlogs	84.23	1.16	0.52	**PlantBlogs**	78.99	2.36	1.06
BotLit	28.19	17.01	7.61	**S800**	86.99	2.85	1.27

Table 11: Average F_1-scores, standard deviation (SD) and standard error (SE) to measure the overall performance variability during 5-fold cross-validation.

After an initial test run on all datasets and an inspection of the model performance over a total of 100 epochs, we found that the scores did not significantly improve during the later epochs.[1] We therefore limited the maximum number of epochs to 50 in order to reduce training time. In the following evaluation tables, the presented scores correspond to the average values computed over the five cross-validation folds. Despite the potential randomness typical for neural models, we report all results with two decimal places to convey the partly subtle differences in performance. Table 12 gives an overview on the best-performing neural models per training corpus.

		German				English		
	Model	P	R	F	**Model**	P	R	F
Combined	**char_dim 29**	94.94	94.98	94.96	**dropout 0.7**	88.54	85.60	87.04
Wiki	**dropout 0.3**	**94.96**	**95.11**	**95.04**	**char_dim 29**	88.65	**86.96**	**87.79**
TB	**dropout 0.7**	90.55	89.11	89.82	**capdim 1**	71.06	54.02	61.12
PlantBlog	**capdim 1**	87.43	83.04	85.07	**char_dim 50**	83.94	76.97	80.22
BotLit/S800	**dropout 0.3**	67.81	36.62	47.21	**dropout 0.7**	**90.83**	84.64	87.60

Table 12: Best-performing neural models per training corpus (*Combined*, *Wiki*, *TB*, *PlantBlog*, *BotLit/S800*).

[1] One epoch represents the iteration over all input training examples, i.e. the sentences in the training dataset [Carreras et al., 2003].

For German, the models trained on the Wikipedia abstracts with a dropout rate of 0.3 and the combined dataset with a character embedding size of 29 achieve the best results. Except for the *BotLit* corpus, which is small in size and contains historical spelling variants of vernacular names, the evaluation showed F_1-scores >89% for the German *TB* and >85% for the *PlantBlog* corpus. Similarly, the English *Wiki* model using a character embedding dimension of 29 outperformed the other datasets with 87.79% in F_1-score, obtaining, however, generally lower F_1-scores as compared to the German dataset. The model trained on the *S800* corpus [Pafilis et al., 2013] using a dropout rate of 0.7 reached a competitive performance with an outstanding precision of >90%, despite being much smaller in size. This is explainable due to the predominant usage of Latin botanical names in this scientific text genre. We hypothesize that the higher performance on the German datasets can be explained due to generally more frequent occurrences of both vernacular and scientific entity mentions (see Table 7 in Section 3.1.3), at least for the *Wiki* and the *TB* corpus.

4.2.1 Baseline

Since there are no multilingual NER systems focusing on the recognition of both vernacular and scientific botanical entities that could be used as a baseline for model comparison, we created a baseline system using the default training parameters provided by Lample et al. [2016]. The default parameters include: no lowercasing of input, a character embedding size of 25, a character LSTM hidden layer size of 25, a bidirectional LSTM for characters, a token embedding dimension of 100, a token LSTM hidden layer size of 100, a bidirectional LSTM for words, no pre-trained embeddings, a conclusive CRF layer, a dropout rate of 0.5, the stochastic gradient descent (SGD) optimization method, a learning rate of 0.01, and gradient clipping [Lample et al., 2016]. For this baseline, we did not integrate any language-specific pre-trained embeddings. Accordingly, all input embeddings were directly trained on the input data with a token embedding dimension of 100. All digits in the input data are represented as zeros to reduce training time (parameter `zeros=True`).

4.2.2 Adding Distributional Information with Word Embeddings

To enrich the input with distributional information [Mikolov et al., 2013], we integrated the pre-trained *FastText* word embeddings with a dimension of 300 [Grave et al., 2018]. These German and English word vectors are trained on Common Crawl and Wikipedia and use a continuous bag of words (CBOW) model architecture, a window size of 5, and character n-grams of length 5 [Grave et al., 2018].

	German			English				German			English		
	P	R	F	P	R	F		P	R	F	P	R	F
Combined dataset							**PlantBlog dataset**						
1st training round							**1st training round**						
baseline	96.26	91.73	93.94	92.71	88.39	90.49	baseline	74.20	66.53	65.80	86.81	61.70	72.08
pre_emb	95.33	95.44	95.38	93.92	90.94	92.40	pre_emb	83.71	79.65	81.52	88.42	70.77	78.54
2nd training round							**2nd training round**						
baseline	95.11	91.45	93.24	88.28	82.35	85.20	baseline	74.96	68.77	70.26	78.96	60.97	68.58
pre_emb	94.88	94.96	94.92	88.60	84.61	86.55	pre_emb	87.08	81.66	84.23	83.16	75.26	78.99

Table 13: Baseline model performance and improvements after adding pre-trained embeddings for combined dataset (left) and *PlantBlog* dataset (right).

As visible in Table 13, integrating the pre-trained *FastText* embeddings [Grave et al., 2018] with a token embedding dimension of 300 resulted in an overall performance improvement as compared to the baseline models.[2] Especially the models trained on the smaller datasets improved notably after adding the pre-trained embeddings. For instance, we can report an increase of +15.72% in F_1-score[3] for the German *PlantBlog* corpus during the first training round and of +13.97% in F_1-score during the second round. We hypothesize that the systematic writing style of the Wikipedia genre, which also constitutes most of the combined dataset, is the reason for the marginal improvements regarding the models trained on the combined datasets (+1.92% in F_1-score for the combined English dataset in the first training round and +1.35% in the second round). Training such in-domain character-level and word-level representations directly on the input data as in the baseline systems appears to model the distributional information for such large datasets fairly well.

4.2.3 Dropout Training

The integration of dropout training is a popular method to improve generalization and prevent overfitting in deep learning with neural networks [Srivastava et al., 2014; Cheng et al., 2017]. To put it briefly, dropout training implies "multiplying neural net activations by random zero-one masks during training" [Cheng et al., 2017]. Thus, units and their connections are randomly dropped to reduce the tendency of co-adaptation between the single units [Srivastava et al., 2014]. The dropout rate `p`

[2] A full overview on the model performance can be found in Appendix B, Table 31.

[3] Please note that the reported improvements in this thesis correspond to percentage points.

represents the probability of the mask value being one [Cheng et al., 2017; Srivastava et al., 2014]. Dropout training additionally ensures that both character-level and token-level representations are used during learning and hence prevents the final model from depending too much on one representation [Lample et al., 2016].

		German			English		
		P	R	F	P	R	F
Combined dataset	pre_emb + drop0.5	**94.88**	**94.96**	**94.92**	**88.60**	84.61	86.55
	pre_emb + drop0.2	94.82	94.80	94.80	88.17	**85.90**	87.02
	pre_emb + drop0.3	**94.88**	94.89	94.88	88.58	84.51	86.49
	pre_emb + drop0.7	94.42	95.18	94.80	88.54	85.60	**87.04**
Wiki	pre_emb + drop0.5	94.75	95.17	94.96	88.12	86.61	87.35
	pre_emb + drop0.2	94.59	**95.37**	94.98	87.68	**86.81**	87.24
	pre_emb + drop0.3	**94.96**	95.11	**95.04**	88.32	85.43	86.84
	pre_emb + drop0.7	94.51	95.04	94.77	**88.39**	86.51	**87.44**
TB	pre_emb + drop0.5	89.77	**89.56**	89.65	63.64	55.26	58.20
	pre_emb + drop0.2	88.36	89.08	88.71	**63.76**	54.44	58.62
	pre_emb + drop0.3	89.57	89.14	89.34	66.60	54.73	60.00
	pre_emb + drop0.7	**90.55**	89.11	**89.82**	68.31	**56.23**	**61.31**
PlantBlog	pre_emb + drop0.5	87.08	81.66	84.23	83.16	75.26	**78.99**
	pre_emb + drop0.2	85.48	82.76	83.88	82.37	**75.63**	78.82
	pre_emb + drop0.3	85.93	**83.36**	**84.53**	81.82	74.87	78.17
	pre_emb + drop0.7	**87.13**	78.99	82.82	**86.16**	71.56	78.05
BotLit/S800	pre_emb + drop0.5	60.09	21.01	28.19	89.59	84.61	86.99
	pre_emb + drop0.2	61.52	35.94	44.43	89.21	83.30	86.13
	pre_emb + drop0.3	**67.81**	**36.62**	**47.21**	90.14	82.98	86.39
	pre_emb + drop0.7	40.00	6.50	11.11	**90.83**	**84.64**	**87.60**

Table 14: Evaluation of dropout training with rates of 0.5 (default), 0.2, 0.3 and 0.7 for all datasets in the second training round.

To improve generalization performance, we experimented with lowered and augmented dropout rates during training. Lample et al. [2016] suggest a dropout rate `p` of 0.2 for English and 0.3 for German. In these experiments, all baselines and *pre_emb* models used a balanced rate (`p` = 0.5). We additionally explored the impact of lowered and augmented rates of 0.2, 0.3 and 0.7 for both languages. As visible in Table 14, using a lowered dropout rate of 0.3 in combination with pre-trained embeddings resulted in improvements for the German *Wiki*, *PlantBlog* and *BotLit* corpora. Contrary to the suggested rate for English by Lample et al. [2016], the largest improvements can be observed when adopting a higher dropout rate of 0.7. In addition, we found that the language-specific adaptation of the dropout rate can be beneficial in cross-corpus settings and improves model generalization across different text genres (see Section 4.3 for a cross-corpus model evluation).

4.2.4 Character Embedding Dimension

As previously mentioned, the neural models do not only rely on token-level distributional information, but also on character-based representations learned directly from the input data. These combined representations capture distributional sensitivity (word embeddings) and orthographic sensitivity (character embeddings) [Lample et al., 2016; Ling et al., 2015].

		German			English		
		P	R	F	P	R	F
combined dataset	pre_emb + char_dim25	94.88	94.96	94.92	**88.60**	84.61	86.55
	pre_emb + char_dim29	**94.94**	94.98	**94.96**	88.32	**85.19**	**86.73**
	pre_emb + char_dim50	94.58	**95.28**	94.93	87.76	85.18	86.44
Wiki	pre_emb + char_dim25	**94.75**	95.17	94.96	88.12	86.61	87.35
	pre_emb + char_dim29	94.60	95.31	94.95	**88.65**	**86.96**	**87.79**
	pre_emb + char_dim50	94.65	**95.32**	**94.98**	88.08	86.34	87.19
TB	pre_emb + char_dim25	89.77	89.56	89.65	63.64	**55.26**	58.20
	pre_emb + char_dim29	**90.04**	89.31	89.65	**66.47**	53.42	**59.18**
	pre_emb + char_dim50	89.75	**89.75**	**89.75**	63.03	54.61	58.31
PlantBlog	pre_emb + char_dim25	87.08	**81.66**	**84.23**	83.16	75.26	78.99
	pre_emb + char_dim29	**87.57**	80.85	83.96	**84.90**	73.43	78.71
	pre_emb + char_dim50	87.51	81.09	84.04	83.94	**76.97**	**80.22**
BotLit/S800	pre_emb + char_dim25	60.09	21.01	28.19	89.59	**84.61**	86.99
	pre_emb + char_dim29	**71.08**	**29.48**	**41.01**	**90.43**	83.93	**87.04**
	pre_emb + char_dim50	55.00	14.96	22.44	89.04	83.43	86.10

Table 15: Evaluation of the character embedding dimension using an embedding size of 25 (default), 29 and 50 for all datasets in the second training round.

The combination of both representations encourages the models to rely on all the available morphological and orthographic evidence that a sequence of tokens is (or is not) a named entity. Especially for morphologically rich languages such as German, we expected these character-level representations to be advantageous. Modeling extensive plant names such as "Unserer-lieben-Frauen-Handschuh" ('purple fox-glove', *Digitalis purpurea*) or "Schmetterlingsblütenartigen" (*Fabales*) could hence yield more accurate representations. To explore the impact of the character embedding dimension parameter, we trained two supplementary models for German and English using an augmented embedding size of 29 and 50 (default = 25). Contrary to our expectations, the evaluation in Table 15 did not reveal a consistent tendency for better performance when using either 29 or 50 as a character embedding dimension. Except for the German *PlantBlog* model, the German models experienced marginal improvements using a character embedding dimension of 29 (combined dataset) and

50 (*Wiki* and *TB* dataset). The largest improvement are observable for the German *BotLit* corpus with +12.82 in F_1-score using an embedding size of 29. For English, we can observe minor improvements using an embedding size of 29 (combined dataset, *Wiki*, *TB* and *S800* dataset) and 50 (*PlantBlog* corpus).

4.2.5 Capitalization Feature Dimension

Capitalization is a valuable orthographic and shape-related feature widely used in feature-engineered NER systems [Yadav and Bethard, 2018]. To explore the potentially beneficial impact, we trained a supplementary model for German and English and included an additional capitalization dimension layer during training.

		German			English		
		P	R	F	P	R	F
combined dataset	pre_emb	**94.88**	**94.96**	**94.92**	**88.60**	84.61	86.55
	pre_emb + capdim	94.20	95.55	94.87	88.58	**84.70**	**86.59**
Wiki	pre_emb	**94.75**	95.17	**94.96**	88.12	**86.61**	**87.35**
	pre_emb + capdim	94.24	**95.71**	**94.96**	**88.22**	85.12	86.63
TB	pre_emb	**89.77**	89.56	**89.65**	63.64	**55.26**	58.20
	pre_emb + capdim	88.97	**89.88**	89.40	**71.06**	54.02	**61.12**
PlantBlog	pre_emb	87.08	81.66	84.23	**83.16**	**75.26**	**78.99**
	pre_emb + capdim	**87.43**	**83.04**	**85.07**	83.04	74.42	78.44
BotLit/S800	pre_emb	60.09	21.01	28.19	89.59	**84.61**	**86.99**
	pre_emb + capdim	**66.59**	**27.82**	**38.24**	**89.84**	83.87	86.71

Table 16: Evaluation of the capitalization feature dimension for all datasets in the second training round.

For this parameter, an array of numerical values represents each input sentence: The value 0 stands for a lowercase word, 1 for all uppercase characters, 2 for words having only the first letter capitalized and 3 for words having any of the other characters in the string capitalized. We expected this parameter to capture entity-specific shape patterns and, thus, to contribute to more accurate tagging results. In Latin species names, for instance, an uppercase genus name usually precedes a lowercase epithet. Similarly, German uppercase adjectives designate descriptive species properties such as in "Kriechender Gänsefuß" ('creeping tormentil', *Potentilla reptans*). The evaluation showed that including the capitalization feature dimension can marginally improve model performance for small German datasets (*PlantBlog* and *BotLit* dataset). In addition, we observed an improvement of +2.92% in F_1-score for the English *TB* corpus and marginal improvements for the combined dataset.

For the other English datasets, no improvements related to the capitalization feature can be reported. We assume that the usually lower-cased English vernacular plant names are the reason for this behavior. Regarding cross-corpus applications and generalization performance on different text genres and unseen entities, the capitalization feature dimension is highly promising (see Section 4.3).

4.2.6 Model Performance Per Entity Label

In the experiments, we adopted a fine-grained set of nine distinct hierarchical entity labels. This includes two distinct classes for German and English vernacular names on the taxonomic levels of species and plant families and seven taxonomic classes for Latin names. Table 17 displays the performance per entity label of the German and English models when using pre-trained embeddings on the combined datasets.

	German					English			
	No. of entities	**P**	**R**	**F**		**No. of entities**	**P**	**R**	**F**
de_species:	4202	90.20	92.26	91.22	**en_species:**	1587	77.82	76.19	77.00
de_fam:	1251	99.52	99.36	99.44	**en_fam:**	196	95.92	94.95	95.43
lat_species:	1580	96.58	98.58	97.57	**lat_species:**	846	93.03	92.81	92.92
lat_genus:	991	95.96	95.48	95.72	**lat_genus:**	854	92.86	85.73	89.15
lat_fam:	1067	99.72	100.00	99.86	**lat_fam:**	483	98.14	99.16	98.65
lat_subfam:	99	98.99	100.00	99.49	**lat_subfam:**	41	97.56	76.92	86.02
lat_class:	23	91.30	100.00	95.45	**lat_class:**	5	80.00	66.67	72.73
lat_order:	46	100.00	97.87	98.92	**lat_order:**	35	97.14	77.27	86.08
lat_phylum:	4	100.00	80.00	88.89	**lat_phylum:**	5	80.00	100.00	88.89
total/average:	9263	96.91	95.95	96.28	**total/average:**	4052	90.27	85.52	87.43

Table 17: Model performance per entity label evaluated on the first fold of the combined dataset for German (left) and English (right).

To answer research question 3 ("How well does the tagger perform with regard to different classes of entities?"), we found that the labels comprising entities with systematic suffix patterns such as *-ceae* for Latin plant families (`lat_fam`), *-oideae* for Latin subfamilies (`lat_subfam`) or *-gewächse*, *-blütler* or *-family* for the vernacular family names (`de_fam` and `en_fam`) are identified reliably in both languages. More heterogeneous categories such as the scientific and vernacular taxonomic level of species (`lat_species`, `de_species`, `en_species`) are often mismatched. Moreover, low-frequent entity labels (`lat_class`, `lat_phylum`) represent a challenge for the neural tagger. The inclusion of additional data containing mentions from these entity classes could be a possible way to augment the recall and to guarantee a consistent performance per entity label. We hypothesize that the performance on

the vernacular species level is higher for German due to the higher concentration of entity occurrences in general (4,202 species mentions for German and 1,587 for English). Presumably, the German introductory Wikipedia abstracts simply include more species names on the vernacular level. For both languages, the performance on the Latin family label is outstanding: We can report an F_1-score of 99.86% for the German combined dataset and an F_1-score of 98.65% for English. The Latin species label (`lat_species`) achieved F_1-scores of >90% for both languages. We assume that the performance is lower for English due to the more heterogeneous Latin species label, including abbreviations such as *B. perennis* for *Bellis perennis*, a phenomenon which is virtually non-existent in the German data.

4.3 Cross-Dataset Evaluation

To explore the domain adaptation potential, we conducted a cross-dataset evaluation. For this purpose, we tested the tagging performance of the models trained on a training fold (80%) of the *Wiki* corpus on a test fold (20%) of the other datasets.[4]

	German				English			
	Model	P	R	F	**Model**	P	R	F
TB								
best-performing Wiki model	**dropout 0.3**	82.40	81.83	82.12	**capdim 1**	68.07	**77.14**	72.32
best-performing TB model	**dropout 0.7**	**90.55**	**89.11**	**89.82**	**dropout 0.7**	**68.31**	56.23	**61.31**
PlantBlog								
best-performing Wiki model	**dropout 0.3**	**92.16**	**83.43**	**87.58**	**chardim 29**	72.65	58.62	64.89
best-performing PlantBlog model	**capdim 1**	87.43	83.04	85.07	**char_dim 50**	**83.94**	**76.97**	**80.22**
BotLit/S800								
best-performing Wiki model	**dropout 0.7**	**100.00**	**45.00**	**62.07**	**pre_emb**	87.80	80.90	84.21
best-performing BotLit/S800 model	**dropout 0.2**	61.52	35.94	44.43	**dropout 0.7**	**90.83**	**84.64**	**87.60**

Table 18: Best-performing models in individual and cross-corpus evaluation.

The results in Table 19 show that low-effort models trained on large Wikipedia datasets can achieve a satisfying performance on a variety of text genres ranging from mountaineering reports (*TB*), blog articles (*PlantBlog*) to botanical literature (*BotLit/S800*). As compared to the results from the individual dataset evaluation

[4]In addition, we analyzed the cross-corpus adaptation capacity of the models trained on the smaller corpora (*TB*, *PlantBlog* and *S800*). Their performance is, however, not comparable to the *Wiki* models. Thus, we did not include them in the cross-corpus evaluation.

(see Table 18), the cross-corpus scores outperformed the single-corpus performance in several cases for German. For the German *PlantBlog* corpus, we observed an increase in F_1-score of +2.51%. The performance on the *BotLit* corpus increased by +17.64% in F_1-score. To answer research question 4 ("Can neural models trained on large, low-effort datasets such as Wikipedia be applied to robustly identify botanical entities in datasets from lower-resourced domains?"), these insights emphasize the potential of large, multilingual Wikipedia models and their application on lower-resourced text genres and domains. For English, none of the cross-corpus Wiki models outperformed the best-performing individual dataset models.

		TB			**PlantBlog**			**BotLit/S800**		
NER-tagging model		P	R	F	P	R	F	P	R	F
Wiki	**pre_emb**	77.03	80.89	78.91	88.46	81.66	84.92	80.00	42.11	55.17
(German)	**chardim 29**	80.38	**81.84**	81.11	93.75	80.36	86.54	**100.00**	**45.00**	**62.07**
	chardim 50	73.73	81.49	77.42	82.35	82.84	82.60	61.54	42.11	50.00
	dropout 0.2	74.56	83.45	78.76	87.97	82.74	85.28	80.00	42.11	55.17
	dropout 0.3	**82.40**	81.83	**82.12**	92.16	**83.43**	**87.58**	80.00	42.11	55.17
	dropout 0.7	85.56	76.53	80.79	**95.00**	78.70	86.08	**100.00**	**45.00**	**62.07**
	capdim 1	76.68	81.40	78.97	90.26	81.76	85.80	80.00	42.11	55.17
Wiki	**pre_emb**	65.85	77.88	71.37	69.35	60.14	64.42	**87.80**	80.90	**84.21**
(English)	**chardim 29**	**69.64**	73.58	71.56	**72.65**	58.62	**64.89**	76.34	83.53	79.78
	chardim 50	64.00	76.19	69.57	70.16	60.00	64.68	79.57	**85.06**	82.22
	dropout 0.2	64.66	71.43	67.87	65.29	56.03	60.31	77.53	81.18	79.31
	dropout 0.3	68.52	70.48	69.48	60.99	60.56	60.78	80.95	77.27	79.07
	dropout 0.7	61.83	**79.41**	69.53	63.24	**61.43**	62.32	76.67	81.18	78.86
	capdim 1	68.07	77.14	**72.32**	64.52	56.34	60.15	84.71	81.82	83.24

Table 19: Evaluation of *Wiki* models on the other datasets (*TB*, *PlantBlog*, *BotLit/S800* in cross-corpus setting.

Lample et al. [2016] underline that dropout training is crucial for good generalization performance. Furthermore, dropout training reduces overfitting, which our system is more prone to because of the systematic annotations. Adapting the dropout rate resulted in robust performance across different text genres for German. The best-performing German Wiki models used a dropout rate of 0.3 for the *TB* and *PlantBlog* dataset and a dropout rate of 0.7 for the *BotLit* corpus. Similar to the insights gained from the single-dataset evaluation, using an augmented dropout rate of 0.7 for English demonstrated a beneficial impact on the performance, especially regarding recall: The Wiki models achieve the highest recall (R) for the English *TB* corpus (R = 79.41) and for the *PlantBlog* corpus (R = 61.43) (see Table 19).

4.4 Tagging Fungi: Evaluation on Unseen Entities

To measure the generalization ability of the *Wiki* models on unseen entities and unseen entity contexts, we conducted an additional cross-corpus experiment. We created two supplementary test sets for German and English from Wikipedia fungi articles by automatically retrieving the introductory abstracts of the articles from the main categories, "Mushroom types" (for English) and "Pilze" ('fungi' for German) using the Wikipedia API. Subsequently, we tested the Wiki model performance on these new datasets containing unseen entities, that is, entities that did not occur in the training data set.

	German	English	Total / Average
domain	mycology (fungi)		
number of tokens	2,986	2,960	5,946
number of types	898	847	1,745
mean token length	5.78	4.69	5.23
number of sentences	163	144	307
mean sentence length	18.31	20.38	19.34

Table 20: Corpus details for supplementary fungi test set.

In addition, we retrieved the abstracts from the subcategories "Agaricales", which correspond to the well-known *gilled mushrooms* including the ubiquitous common mushroom or the poisonous *fly agaric* and the "Boletales" category including delicious boletes like *penny buns*, *birch boletes* or the *dotted stem boletes* [Brandon et al., 2007; Wikipedia, 2018a,b]. Since these two orders contain some of the most popular and wide-spread types of edible and poisonous mushrooms, our intuition was that the introductory Wikipedia abstracts do not only comprise scientific, but also vernacular fungal names. Table 20 depicts the fungus corpus details for German and English. In terms of size, we collected approximately 3,000 tokens per language, which is similar to the size of the *PlantBlog* test set. It should be mentioned, that besides vernacular and scientific fungal names, this test sets also include a few plant names. This contextual co-occurence is mainly due to symbiotic or parasitic relationships between fungi and trees (see Example 1 below) or phylogenetically related lichens (2). In the fungi test set, we annotated both types of entities on the species level, without any additional distinction between different kingdoms or symbiotic relationships (e.g. plants or lichens versus fungi).

1. It was once thought to be mycorrrhizal with *Pinus sylvestris*.

2. *Basidiolichens* are lichenized members of the Basidiomycota, a much smaller group of *lichens* than the far more common *ascolichens* in the Ascomycota.

In total, the English fungi test set comprises 182 vernacular species names, 122 Latin species, 44 generic names, 4 subfamilies, 31 families, 24 orders and 4 classes. The German set includes 229 vernacular species and 62 family names, 192 Latin species, 32 genera, 1 subfamily, 33 families, 6 orders and 1 class name. Table 21 reports the results from applying the *Wiki* models on the language-specific fungi test sets.

	Fungi testset (German)				**Fungi testset (English)**			
Wiki-models	A	P	R	F	A	P	R	F
pre_emb	87.11	83.54	36.40	50.70	94.07	89.42	62.82	73.80
chardim 29	87.31	82.79	37.48	51.60	93.32	87.21	57.99	69.66
chardim 50	87.41	84.65	37.57	52.04	93.97	89.01	62.31	73.30
dropout 0.2	88.18	85.28	41.85	56.15	93.46	86.92	58.85	70.19
dropout 0.3	85.70	80.65	27.73	41.27	93.15	88.16	55.67	68.25
dropout 0.7	86.27	84.50	30.84	45.19	93.76	87.23	61.76	72.31
capdim 1	**88.31**	**86.89**	**42.49**	**57.07**	**94.17**	**89.78**	**63.24**	**74.21**

Table 21: A (accuracy), P (precision), R (recall) and F (F_1-score) for Wiki model performance on unseen entities (fungi test set).

Interestingly, the *capdim* models using an additional layer for the capitalization feature dimension, were particularly successful in this cross-corpus setting dealing with unseen mycological entities. The results for the German fungi test set achieved an F_1-score of maximally 57.07%, while the English *Wiki* models reached a performance of 74.21%. Contrary to the beneficial effect of dropout training mentioned in Section 4.3, the `cap_dim` models using a balanced dropout rate of 0.5 were the most promising parameter combination in this setting. A brief inspection of the tagged output produced by the neural tagger showed that the main source of error is not only concentrated on the level of vernacular names, but also affects scientific genus and species names. As visible in Example 1, the German fungus "Sumpf-Saftling" ('waxcap mushroom') has been correctly identified as a vernacular species name.

1. Der__O **Sumpf-Saftling**__B-de_species (__O **Hygrocybe**__O **helobia**__O [...]

2. **Gyroporus**__B-lat_species **castaneus**__I-lat__species ,__O or__O commonly__O the__O **chestnut**__B-en_species **bolete**__O ,__O is__O a__O small__O ,__O white__O -__O pored__O relation__O of__O the__O **Boletus**__O mushrooms__O

Although Latin fungal species names follow the systematic pattern of usually two tokens with typical, case-congruent suffixes, the name "Hygrocybe helobia" has not been recognized by the neural tagger. We assume that the presence of the form "-cybe", which is typical for mycological names describing a special cap shape [Cundall, 1998], is the reason for this behavior. Example 2 shows that typical and obviously congruent Latin species suffixes (*-us*, *-us*) led to the correct identification of "Gyroporus castaneus". The name "chestnut bolete", conversely, has only been partially recognized by the system. Presumably, the subsequence "chestnut", which also refers to multiple deciduous trees from the genus *Castanea*, is an interfering factor in this case. Finally, the genus name "Boletus" in Example 2 is missed by the tagger despite the presence of the typical Latin suffix *-us*.

To sum up, we observed a particularly beneficial impact of the capitalization feature dimension when tagging unseen entities in the related sub-domain of mycology. Again, the potential of large *Wiki* models for the application on new, potentially lower-resourced genres and sub-domains is worth highlighting.

4.5 Comparison to In-Domain Systems

Since previous plant name recognition systems focus, to the best of our knowledge, on the identification of Latin taxonomic entities in text, we compared the performance of the neural models to our own baseline (Sections 4.2.1 and 4.2.2). In terms of direct comparability, this approach guarantees informative and fair values on an identical set of entity types. Nevertheless, we include a brief overview and evaluation of related in-domain systems in Table 22, regardless of whether they are rule-based [Koning et al., 2005; Sautter et al., 2006], dictionary-based [Gerner et al., 2010; Pafilis et al., 2013; Leary, 2014], or based on machine learning techniques [Akella et al., 2012; Habibi et al., 2017]. All the listed approaches have a strong focus on automatically identifying scientific plant names in text corpora. Except for the *TaxonFinder* developed by Leary [2014], the tools do not distinguish between multiple hierarchical levels in their entity label set. Commonly, they only adopt a single label for "organism" [Akella et al., 2012], "species" [Gerner et al., 2010; Habibi et al., 2017] or "taxon name" [Sautter et al., 2006]. Even if the recognition of vernacular plant names is, at least to some extent, addressed as for example in the *Species* tagger presented by Pafilis et al. [2013], the authors did not conduct an entity-specific evaluation in their publication. Hence, the system's performance regarding named entities on a vernacular level cannot be sufficiently assessed and compared to our approach. In addition, not all of these systems aim for the identification of entities

on different taxonomic levels (species, genus, family, subfamily, etc.). The systems introduced by Habibi et al. [2017] and Gerner et al. [2010], for instance, focus on the identification of taxa on the species level and thus neglect the classification of entities on other taxonomic levels. Especially when dealing with the automatic population and extension of hierarchical knowledge bases, this taxonomic information is, in our opinion, highly important.

System	Method	Languages	Entities (no. of entity labels)	Test Corpus	Authors	Evaluation Results		
						P	R	F
Our System	*bi-LSTM-CRF*	*en (Wiki model)*	*scientific (7), vernacular (2)*	*S800*	*Meraner (2019)*	*87.8*	*80.9*	*84.2*
Our System	*bi-LSTM-CRF*	*en (S800 model)*	*scientific (7), vernacular (2)*	*S800*	*Meraner (2019)*	*90.8*	*84.6*	*87.6*
Biomed. NER	LSTM-CRF	en	scientific (1) (focus on species level)	S800	Habibi et al. (2017)	80.8	87.6	83.6
NetiNeti	probabilistic ML classifier	en	scientific (1) (all taxonomic levels)	ASB	Akella et al. (2012)	98.9	70.5	82.3
gnparser	statistical parsing, CFG	en	scientific (parsing of complex entities, incl. authorship)	1000 name-strings	Mozzherin et al. (2017)	98.9	100	99.4
Species	dictionary-based	en	scientific (1) all taxa, (partially) vernacular	S800	Pafilis et al. (2013)	83.9	72.6	77.8
Linneaus	dictionary-based	en	scientific (1) (focus on species level)	S800	Gerner et al. (2010)	84.3	75.4	79.6
TaxonFinder	dictionary-based	en	scientific (8) (all taxonomic levels)	ASB	Leary et al. (2014)	97.5	54.3	69.7
TaxonGrab	rule-based	en, es, de, fr	scientific (1) (all taxonomic levels)	C.1932	Koning et al. (2005)	96.0	94.0	95.0
FAT	rule-based	en	scientific (1) (all taxonomic levels)	C.1932	Sautter et al. (2007)	92.7	87.8	95.0

Table 22: Comparison to in-domain botanical entity recognition systems.

It should be mentioned, that some of the systems also focus on the identification of species names from other kingdoms (i.e. fungi, animals) [Habibi et al., 2017]. Regrettably, the authors did not present an individual performance evaluation per sub-domain (botany, mycology or zoology). We compare our approach to three of the listed systems, since the authors used the *S800* test set [Pafilis et al., 2013] for their evaluation. Habibi et al. [2017] for example report an F_1-score of 83.6% for the species label, which, unlike our model, also includes the recognition of other organism names. As visible in Table 22, our approach achieved an F_1-score of 87.6% for the model trained with a dropout rate of 0.7 on a subset of the *S800* corpus including only botanical abstracts. We achieved an F_1-score of 84.2% in the cross-corpus evaluation setting using the model trained on Wikipedia abstracts, pre-trained embeddings and a balanced dropout rate of 0.5. Our system also outperformed two other systems that have been evaluated on the *S800* test set: Pafilis et al. [2013] report an F_1-score of 77.8%, while Gerner et al. [2010] achieve a slightly higher F_1-score of 79.6%.

In summary, the comparison to in-domain systems suggests that the neural models presented in this work achieve competitive performance in single-corpus (*S800* model) and cross-corpus (*Wiki* model) settings. Concerning the level of granularity, we want to stress that the adoption of fine-grained, hierarchical entity labels can be beneficial for a variety of tasks such as automatic taxonomy learning and knowledge base population. Our approach shows that an entity label set of nine distinct taxonomic classes can be robustly identified across manifold genres without sacrificing the overall performance.

4.6 Error Analysis

For a better understanding of potential sources of errors and deviations from our initial expectations, we conduct an error analysis in this section. We focus on the following main critical components and their potential interference with the model performance:

1. Preprocessing as a source of error
2. Shape and heterogeneity of botanical entities
3. Language-specific entity ambiguity

4.6.1 Source of Error I: Preprocessing

After the first experiment run and evaluation, we noticed that a decisive factor leading to errors in the data and biased results can be localized upstream in the preprocessing pipeline (see Section 3.2). For instance, the tokenization tools integrated in *NLTK* [Loper and Bird, 2002] and *spaCy* [Honnibal and Montani, 2017] were unable to deal with domain-specific abbreviations. The creation and application of optimized, domain-specific tokenizers could hence result in a higher quality of the preprocessing output. Current approaches in this field combine domain-adapted regular expressions and machine learning for the split-join classification task [Barrett and Weber-Jahnke, 2011]. The adaptation and tailoring of such preprocessing tools was, however, not the main scope of the current project. To avoid erroneous sentence segmentation and truncated Latin plant names, which often contain such botanical abbreviations, we applied a regular expression in order to re-merge sentences in the IOB-annotated files (see Table 23).

regular expression	\n(var|convar|agg|ssp|sp|subsp|x|L|auct|comb|illeg|cv|emend|al|f|hort|nm|nom|ambig|cons|dub|superfl|inval|nov|nud|rej|nec|nothosubsp|p|hyb|syn|synon)(\t.*?)\n\.\t.*?\n
replace pattern	\n\1\.\2\n

Table 23: Find and replace pattern to re-merge sentences that have been split at botanical abbreviations (list after Meades [2018]) during preprocessing.

The correction of these erroneously split sentences allowed us to automatically annotate more complete and extensive plant names using our dictionary-based annotation system (see Table 24). The example below shows that our dictionary-based system was only able to annotate the extensive Latin infraspecies name "Diospyros kaki var. sylvestris" in the second annotation round (right). Previously, the entity mention has been split and truncated during tokenization due to the presence of the botanical abbreviation "var." standing for a species variety. While in the first version of the *Wiki* dataset 7,659 entities have been annotated with the IOB tag `I_lat_species` (inside of a Latin species name), we found 8,309 internal species names in the second annotation round.

1st annotation round				**2nd annotation round**			
A	a	DET	O	A	a	DET	O
variety	variety	NOUN	O	variety	variety	NOUN	O
is	be	VERB	O	is	be	VERB	O
Diospyros	**diospyros**	**PROPN**	**B-lat_species**	**Diospyros**	**diospyros**	**PROPN**	**B-lat_species**
kaki	**kaki**	**NOUN**	**I-lat_species**	**kaki**	**kaki**	**NOUN**	**I-lat_species**
var	var	NOUN	O	**var.**	**var**	**NOUN**	**I-lat_species**
.	.	PUNCT	O	**sylvestris**	**sylvestris**	**PROPN**	**I-lat_species**
				Makino	makino	PROPN	O
sylvestris	sylvestris	PROPN	O	.	.	PUNCT	O
Makino	makino	PROPN	O				
.	.	PUNCT	O				

Table 24: Initial sentence segmentation at botanical abbreviations (left) and corrected version after second annotation round (right).

4.6.2 Source of Error II: Entity Shape and Heterogeneity

Especially for German, the increased variety in shape and length of the annotated entities during the second training run resulted in lower evaluation scores on the silver standard as compared to the first run (see Table 25). The final training

material for both languages contained more variegated surface patterns. These comprise hyphenated compounds and plant names of variable length, ranging from unigrams such as "sage" (Salvia officinalis) to n-grams as in "Jack-go-to-bed-at-noon" (Tragopodon pratensis) for the entity label `en_species`. Moreover, the Latin species label (`lat_species`) presents a larger variability with regard to the number of tokens involved: We may encounter bi-grams such as "Nigella sativa" and four-grams referring to infraspecies names such as "Lactuca sativa var. crispa". As a result, better annotations led to more heterogeneous entity candidates per label and, presumably, to some sort of "confusion" of the LSTM-CRF tagger. Nonetheless, the systematic, semi-automatic corrections based on part-of-speech (POS) patterns or sentence re-merging had a positive impact on the performance as measured by the gold standard, - even if this led to more heterogeneous entities in the first place. To demonstrate this, we evaluated the neural model performance during both stages on the automatically annotated silver standard and on the manually corrected gold standard (see Table 25). This shows the quality of a system trained on silver-labeled data when evaluated on gold annotations.

	German								English							
	silver standard				gold standard				silver standard				gold standard			
	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
1st round																
baseline	98.58	96.52	91.32	93.85	94.13	82.83	79.49	81.12	98.01	87.56	78.31	82.68	97.16	89.99	75.54	82.13
pre_emb	98.92	95.75	94.70	95.22	94.53	84.03	80.86	82.42	98.42	88.94	83.70	86.24	97.27	88.36	78.50	83.14
2nd round																
baseline	98.30	95.38	90.96	93.12	97.61	96.23	88.72	92.32	98.07	88.28	82.35	85.20	98.05	89.91	87.22	88.55
pre_emb	98.81	94.41	95.68	95.04	98.15	96.68	91.79	94.17	98.26	88.6	84.61	86.55	98.17	90.20	88.46	89.32

Table 25: Performance of combined data models on automatically annotated silver standard and manually corrected gold standard over two training rounds.

Table 25 shows that, at a first glance, the performance of the German models on the silver standard decreased from the first to the second evaluation round. For the German combined dataset the F_1-scores fell by -0.73% for the `baseline` model and by -0.18 for the `pre_emb` model. Yet, if looking at the associated performance on the gold standard on the right side, it is visible, that the performance gap between silver standard and gold standard significantly decreased in the second training round. The predictions made by the neural tagger are thus closer to the ground truth, even though the silver standard test sets revealed minor performance decreases. As previously mentioned, we assume that this decrease is due to more heterogeneous German species entities ranging from unigrams (1 single token constituting an entity) to multiword names (entities spanning over multiple tokens).

For English, the performance on the gold standard even outperforms the silver standard scores after the second training round (+3.35% in F_1-score for the English `baseline` and +2.77% for the `pre_emb` model). When considering the performance on the English gold standard as opposed to the silver standard during the first training round, we only observed a marginal decrease of -0.55% for the baseline and of -3.10% for the `pre_emb` model. This contrasts with the performance gap between silver standard and gold standard observable during the first training round for the German data (-12.73% for the `baseline` and -12.80% for the `pre_emb` model).

On the whole, we obtain a model performance of F_1-scores >94% on both silver standard and gold standard for the combined German model in the second evaluation round. For English, we achieve >89% F_1-score performance on the manually labeled gold standard in the second evaluation round.

4.6.3 Source of Error III: Language-Specific Entity Ambiguity

As previously mentioned, language-specific entity ambiguity represents a considerable challenge for the dictionary-based annotation system. Thus, we applied semi-automatic corrections in the second annotation round to obtain more correctly labeled training examples. For instance, we semi-automatically corrected the German cases "Sonnenwende" (1. 'solstice', 2. 'European turn-sole' (*Heliotropium europaeum*)) using regular expressions or context-dependent individual expert decisions. Other corrected ambiguous cases were the German plant names "Winde" (1. 'winds', 2. 'bindweed' (*Convulvulus*)), "Buchs" (1. 'of the book' (genitive), 2. 'Buchs' (location in Switzerland), 3. 'boxwood' (*Buxus sempervirens*)), "Edelweiss" or "Edelweiß" (1. name for guest houses in German-speaking alpine regions, 2. Alpine edelweiss (*Leontopodium nivale*)). Other interesting cases triggering language-specific and context-dependent ambiguity are the German compounds "Berglein" (literally 'small mountain', *Thesium bavarum*) and "Zwerglein" (literally 'little gnomes', *Radiola linoides*). Both vernacular names contain the diminutive suffix *-lein* ('small'). In this case, however, this suffix actually refers to the related genus *Linum* called "Lein" in German ('flax').

For English, we corrected ambiguous color names that, depending on the context, might also refer to plants: "rose" (1. 'rose color', 2. 'species of the genus *Rosa*'), "orange" (1. 'orange color', 2. 'the species *Citrus x sinensis*'), "mauve" (1. 'mauve color', 2. 'species of the genus *Malva*'). The same applies to ambiguous Latin genus names that frequently coincide with female first names or toponyms: "Victoria" (1. first name, 2. world-wide distributed toponym, 3. plant genus from

the *Nymphaeaceae* family), "Rosa" (1. female first name, 2. plant genus from the *Rosaceae* family). An inspection of the tagged output revealed the color "rose" has been correctly labeled as outside (O) after the second training round. We do, however, find one false positive instance referring to "Rose" as a proper name.[5]

4.7 Cross-Lingual Comparison

The shape and quality of language-specific vernacular plant names is highly dependent on ethno-cultural factors and the language-community's individual relationship with certain wild, medical or domesticated plants. Vernacular names for exotic species are often loan words or translations from the native language(s) of those countries where the plant originally grows and, hence, a single designation is mainly consistent across multiple languages (see Examples 1 and 2 in Table 26).[6]

	LAT	EN	DE	FR	ES	IT
1.	**Ravenala madagascariensis**	traveller's tree	Baum der Reisenden	arbre du voyageur	àrbol del viajero	albero del viaggiatore
2.	**Citrus medica var. sarcodactylis**	Buddha's hand	Buddhas-Hand-Zitrone	main de Bouddha	mano de buda	mano di Buddha
3.	**Taraxacum officinale**	dandelion blowball faceclock tell-time bitterwort clockflower swine's snout Irish daisy	Löwenzahn Butterblume Pusteblume Kettenblume Kuhblume Maiblume Pfaffenöhrlein Pferdeblume	dent-de-lion pissenlit cramaillots fausse chicorée liondent salade de taupe laitue des chiens dent de chien	diente(s) de león panaderos ásteres achicoria amarga meacamas amargòn radicheta botón de oro	tarassaco dente di leone dente di cane soffione pisciacane ingrassaporci cicoria asinina grugno di porco
4.	**Rumex alpinus**	alpine dock munk's rhubarb mountain-rhubarb	Alpenampfer Scheißplätschen Butterplätschen Sauplotschen Bergrhabarber Blacken (CH)	oseille des Alpes patience des Alpes rapponti rhubarbe des moines rumex des Aples	romaza alpina vinagretas vinagreras	rabarbaro alpino romice alpino erba pazienza acetosa rubice alpino lacasso

Table 26: Cross-lingual comparison of selected exotic and autochthonous vernacular species names for English, German, French, Spanish and Italian.

Conversely, autochthonous plants with a long tradition as either medical, shamanic, or food plants are given manifold creative names, usually reflecting the specific scope of application, active substances, habitat, or therapeutical effect (see Examples 3 and 4). The German "Warzenkraut" ('nipplewort') reflects the beneficial effect of *Chelidonium majus*[7] for the successful treatment of warts [Achmüller, 2016]. Ex-

[5]Sentence context: Rose considers his name and nature uncertain [...]. (*Wiki* corpus)

[6]Examples: Pfeiffer [1898]; Achmüller [2016]; GBIF [2018]; Roskov et al. [2018]; Wikipedia [2018c].

[7]In some dialects, this vernacular name may also refer to species of the genus *Euphorbium*.

ample 3 shows that vernacular names may also be subject of linguistic variation and phonetic adaptation processes for loan words: The English common name for *Taraxacum officinale* "dandelion" constitutes a contraction of the French expression "dent-de-lion" ('lion's tooth') [Cresswell, 2010]. Other names for this species such as "tell-time" refer to its folk usage of telling time:

> "The dandelion is called the rustic oracle;
> its flowers always open about 5 A.M. and shut at 8 P.M.,
> serving the shepherd for a clock." [Chamberlain, 1896]

What is more, vernacular expressions such as "blowball", the German "Pusteblume" or the Italian "soffione" refer to the custom of blowing the mature seeds from a dandelion's globe. Interestingly, the Italian form "dente di leone" ('lion's tooth') has a related variant, namely "dente di cane" ('dog's tooth'). The same is true for the French expression "dent de chien" ('dog's tooth'), in German, on the other hand, no equivalent expression can be found for *Taraxacum*. The interesting case of *Rumex alpinus* commonly referred to as "Scheißplätschen" ('shit leaves') or "Butterplätschen" ('butter leaves') in different German dialects from Southern Germany, Austria, Switzerland and South Tyrol conveys two distinct aspects of information: According to Achmüller [2016], the expression "Scheißplätschen" is related to the typical habitat of the plant: Alpine meadows covered in cowpats. The second expression "Butterplätschen" is linked to the fact that farmers in Alpine regions used to wrap up fresh butter in the leaves to extend its storage life [Achmüller, 2016].

But how do such linguistic factors such as the level of structural complexity, i.e. hyphenated compounds or names spanning over multiple tokens, influence the performance of the LSTM-CRF models on the test data for German and English? For both languages, multiword expressions (MWE) combining an adjective describing the species and a subsequent noun referring to the plant genus occur frequently (see Examples 1 and 2 in Table 27). With regards to the phenomenon of compounding, English names are split at hyphens, e.g. "lime prickly-ash" (*Zanthoxylum fagara*) (see Example 3), whereas German compounds are represented as single tokens such as "Mücken-Händelwurz" (*Gymnadenia conopsea*) in Example 4. An inspection shows that, in most cases, such instances are predicted reliably by the LSTM-CRF-tagger, even if the silver standard annotations are missing or incomplete (see Examples 1, 3 and 4).

With regard to research question 2 ("How do linguistic phenomena such as compounding influence the performance of the neural NER models?"), we can say that even though linguistic phenomena such as compounding and hyphenated multiword

	Structure	Token	Lemma	POS	IOB (silver)	IOB (predicted)
1.	**MWE**	Madagascan	madagascan	ADJ	O	**B-en_species**
	(English)	bamboo	bamboo	NOUN	B-en_species	**I-en_species**
2.	**MWE**	Fünfblättriger	zweifelhaft	NN	B-de_species	**B-de_species**
	(German)	Wilder	wild	ADJA	I-de_species	**I-de_species**
		Wein	Wein	NN	I-de_species	**I-de_species**
3.	**hyphenated**	lime	lime	NOUN	B-en_species	**B-en_species**
	compound	prickly	prickly	AJD	I-en_species	**I-en_species**
	(English)	-	-	PUNCT	O	**I-en_species**
		ash	ash	NOUN	O	**I-en_species**
4.	**hyphenated compound (German)**	Mücken-Händelwurz	<unk>	NN	O	**B-de_species**

Table 27: Cross-lingual tagging comparison of structurally complex vernacular entities (compounds, multiword expressions) for German and English.

expressions represent a major challenge for domain-specific named entity recognition (NER), robust performance is achievable using neural, language-specific NER models trained on high-quality silver labels.

4.8 Summary and Discussion

Concerning research question 1 ("How well does the state-of-the-art bidirectional LSTM-CRF architecture for named entity recognition perform on domain-specific scientific and non-scientific text genres?"), the model evaluation presented in this chapter showed that both scientific and vernacular plant names can be identified and hierarchically classified across multiple text genres and languages. Especially for smaller datasets such as *TB*, *PlantBlog* and *BotLit/S800*, we observed considerable improvements after integrating the pre-trained *FastText* word embeddings with an embedding size of 300 [Grave et al., 2018]. Due to time constraints, we did not focus on the comparison of different types of embeddings. Presumably, the application of domain-specific (i.e. trained on PubMed abstracts for English [Habibi et al., 2017]), context-sensitive embeddings [Akbik et al., 2018], or deep contextualized word representations [Peters et al., 2018] could yield more fine-grained and

accurate word representations for the detection and classification of named entities in this specific domain. To avoid extensive semi-automatic post-corrections of the data resources, the development of flexible domain-adaptable preprocessing tools including tokenizers and part-of-speech-taggers could be profitable.

Contrary to our expectations, an augmented character embedding dimension resulted in only marginal improvements. Regarding the performance per entity label, we observed F_1-scores of >98% (English) and of >99% (German) for the entity label `B-lat_fam` having the systematic suffix *-ceae*. Including the capitalization feature dimension as an additional layer during training proved to be particularly beneficial in cross-corpus settings dealing with unseen entities from related domains such as mycology (Section 4.4). Despite minor decreases in performance on the silver-labeled data in the second training round, we observed F_1-scores of >94% for the German combined dataset on the gold standard test set. For the English combined dataset, we can report a maximum performance of 89% F_1-score on the gold standard data. Lastly, we want to highlight the potential of low-effort neural models trained on Wikipedia articles for processing lower-resourced text genres such as historical botanical literature and new domains as demonstrated in the cross-corpus evaluation setting (Sections 4.3 and 4.4). Our error analysis and cross-lingual entity comparison outlined potential pitfalls due to the versatile shapes and structural patterns of vernacular plant names, which represent a major challenge when tagging plants in multilingual texts.

5 Linking Plants: Botanical Entity Linking and Visualization

The final step in our approach involves the disambiguation of the candidate entities proposed by the best-performing German or English tagging model and the subsequent linking to an entry in a knowledge base. We used the *Catalogue of Life* (CoL) [Roskov et al., 2018] as a reference database to link the proposed scientific and vernacular candidates to a unique identifier. This international resource provides a public webservice to retrieve any catalogued organism name from the database and offers JSON and XML output formats to return the results. Despite being the most comprehensive global database for international species from different kingdoms such as plants, fungi or animals, it is not yet complete. As a matter of fact, the *CoL* partially lacks, especially for German, historical and current vernacular names for several species. To overcome this obstacle, we used a language-specific lookup table that we created by combining several structured resources retrieved during the data collection phase.

In the following, we discuss the outcomes and insights gained from the final entity linking stage. We analyze the overall coverage of the botanical entity linking when using an international reference knowledge base for disambiguation. In particular, we aim to assess whether or not the majority of the entity candidates can be directly linked to an entry in the database. If direct linking fails, can the integration of a customized lookup table help to overcome this hurdle?

5.1 Querying Botanical Reference Databases

Querying the *Catalogue of Life* (CoL) webservice for botanical entities returns a wide variety of valuable taxonomic information. For this project, we chose the JSON output format and retrieved the entity's database ID, the taxonomic rank (species, genus, order, class, etc.) of the queried species, the official name status according to the *CoL* (accepted name, synonym or common name), a URL to the

corresponding database page, and a currently accepted scientific name. The latter is the key information for disambiguating synonyms, outdated variants or vernacular names. To illustrate this, querying the outdated, synonymous name for the *lemon verbena*, "Aloysia triphylla", returns the name status "synonym" and the associated, currently accepted name variant, namely *Aloysia citriodora*:

Query:

```
http://webservice.catalogueoflife.org/col/webservice?name=Aloysia+triphylla&format=json
```

Result:


```
"results": [
  {
    "id": "3a05ac98506a8ef1a153537f1360adb3",
    "name": "Aloysia triphylla",
    "rank": "Species",
    "name_status": "synonym",
    "online_resource": "",
    "source_database": "WCSP: World Checklist of Selected Plant Families",
    "source_database_url": "http://apps.kew.org/wcsp/",
    "url":
    "http://www.catalogueoflife.org/col/details/species/id/3834e4a9a757a29d7
    b3",
    "accepted_name": {
      "id": "3834e4a9a757a29d7ffc9b99c18972bf",
      "name": "Aloysia citrodora",
      "rank": "Species",
      "name_status": "accepted name",
```


Figure 5: Catalogue of Life query result in JSON-format for *Aloysia triphylla*.

While candidates in the Linnaean binomial format (genus name + species epithet) can be robustly retrieved and linked to the botanical reference database, finding and disambiguating alternative name forms such as outdated synonyms, abbreviations, vernacular names, or names following other nomenclatures, holds a variety of challenges. As visualized in Figure 5, we addressed this issue by automatically checking the specific name status of each entity candidate. This allowed us not only to automatically detect alternative name forms, but also to disambiguate and link such cases to a currently accepted scientific name.

With regard to vernacular name entries, the *CoL* database is, however, only partially complete and reliable, especially when it comes to the retrieval of non-English common names. For this reason, we reused parts of the tabular data provided by several institutions or private experts for the initial creation of the gazetteers. In so doing, we were able to build a customized lookup table containing vernacular names and their associated scientific equivalent(s) (see Table 28). In terms of size, the German lookup table comprises 38,831, the English table 112,406 entries. A ver-

nacular name might be linked to multiple scientific names, which is explainable due to multiple factors: a vernacular name might refer to different taxa, e.g. "hen-and-chicks" points to species from *Jovibarba* or *Sempervivum*. Alternatively, differing degrees of structural complexity for the associated Latin names can lead to multiple entries per vernacular name. This involves infraspecies names such as varieties (var.) or subspecies (ssp.), authorship information including a botanist's name or alternative name forms. For example, the entry for the German name "Zucchini" ('green squash, zucchini') points to the main species name *Cucurbita pepo* and to the infraspecies name *Cucurbita pepo subsp. pepo*. To avoid querying extensive scientific names including authorship information in the reference database, we chose the binomial species name.

Vernacular Name	**Latin Name(s)**
Aargauer Gold-Hahnenfuß	Ranunculus argoviensis
Abelie	Abelia chinensis
Abelie	Abelia x grandiflora
Abendländischer Lebensbaum	Thuja occidentalis
Abendländischer Lebensbaum	Thuja occidentalis 'Smaragd'
abessinischer Kohl	Brassica carinata
abessinischer Meerkohl	Crambe hispanica
abessinischer Meerkohl	Crambe hispanica subsp. abyssinica
Abgebiss-Pippau	Crepis praemorsa
Abgebissener Pippau	Crepis praemorsa
Abgebissener Pippau	Crepis praemorsa (L.) Tausch
Abgebissener Pippau	Crepis praemorsa (L.) Walther
Akanthusgewächse	Acanthaceae

Table 28: Lookup table for vernacular and associated scientific names.

5.2 Entity Linking Performance

To assess the overall entity linking coverage, we used the four annotated corpora and measured how many entities can be linked directly and how many require additional lookups before being linked. Table 29 shows that the task of linking domain-specific entity candidates to a unique database entry is often hindered by missing database entries. This is particularly the case for vernacular named entities. Our evaluation on the four language-specific datasets shows that the integration of a lookup table for mapping vernacular names to an associated Latin equivalent, is especially beneficial for the German data.

The linking coverage increases by approximately +8% (for *Wiki* and *TB* dataset) to +16% (for the *PlantBlog* dataset) for German, whereas it only increases by +0.8% (*Wiki* dataset) to maximally +7% (*PlantBlog* dataset) for the English corpora.

	Wiki		**TB**		**PlantBlog**		**BotLit**	
	German	English	German	English	German	English	German	English
total no. of sentences	13,882	15,280	2,289	153	720	876	42	960
total entity candidates	42,490	19,457	3,268	398	843	685	84	339
linked entities (direct)	8,568 (20.16%)	8,387 (43.11%)	613 (18.76%)	268 (67.33%)	267 (31.67%)	227 (33.13%)	35 (41.67%)	86 (25.37%)
linked entities (lookup table)	11,773 (27.71%)	8,540 (43.89%)	859 (26.29%)	274 (68.84%)	390 (46.26%)	246 (35.91%)	43 (51.19%)	91 (26.84%)
linked entities (lookup table+lower-casing)	12,031 (28.31%)	8,540 (43.89%)	862 (26.38%)	280 (70.35%)	393 (46.62%)	275 (40.15%)	43 (51.19%)	93 (27.43%)

Table 29: Entity linking performance on the four datasets.

Regarding research question 5 ("Can entity linking be applied for the semantic disambiguation of both scientific and vernacular plant names?"), these outcomes suggest that the linking performance is not yet satisfying, especially on the level of vernacular names for non-international languages such as German. The integration of comprehensive lookup tables can be an acceptable workaround for the time being.

In the case of multiple database instances returned by the API request, we chose the response's first instance which corresponds, at least for scientific names, to the most relevant or the most widely accepted name variant [Roskov et al., 2018]. For example, the API call for the *black caraway* (`name=Nigella+sativa`) returns three alternative instances from the *CoL* database: 1. Nigella sativa, 2. Nigella sativa var. hispidula, 3. Nigella sativa var. brachyloba. Nevertheless, we encountered false positive linking results, where the first returned instance did not correspond to the entity we were looking for.

Frequently, the returned instances were fuzzy matches, which actually referred to a fungal or viral disease that often occurs on this exact plant (see Examples 1, 2 in Table 30). This behavior of the *CoL* webservice is unexpected, especially because we did not use any wildcards in the query expression to additionally retrieve more extensive matches. Moreover, semantically vague vernacular names such as "vine" (referring to different species of climbing plants) are often linked to untypical scientific names (3). In this example, we would rather expect a more prototypical climbing plant, e.g. the grapevine (*Vitis vinifera*) instead of *Adlumia fungosa*,

	Query	Gold disambiguation	Linking result (*CoL* webservice)
1.	***cucumber***	Cucumis sativus ('cucumber')	Aureusvirus: Cucumber leaf spot virus
2.	***Oryza sativa***	Oryza sativa ('Asian rice')	Endornavirus: Oryza sativa endornavirus
3.	***vine***	Vitis vinifera ('grapevine')	Adlumia fungosa

Table 30: False positive entity linking examples for *cucumber*, *Asian rice* and *vine*.

which is commonly known as *Allegheny vine*.[1] In addition, we found that linking abbreviated forms of Latin names resulted in some difficulties. Since the full plant name might not always co-occur within the same sentence, it is not always possible to disambiguate the abbreviated form before querying the database. As a result, querying the abbreviated form "L. japonicus" referring to the wild legume *Lotus japonicus* using the query expression `name=L.+japonicus` returns an empty JSON-object. The *CoL* webservice does only allow the use of wildcards at the very end of a query expression and not at the inside of the query [Roskov et al., 2018], which could have been a possible workaround to find database entries for these abbreviated names. Thus, these abbreviated forms should be treated with caution. We would like to emphasize that the disambiguation of such abbreviated forms without contextual information on the document-level, is, in such cases, not always possible. The abbreviated name "N. aquatica", for instance, can refer to the water tupelo *Nyssa aquatica* or to the North American lake cress *Neobeckia aquatica*. Pafilis et al. [2013] successfully address this issue in their dictionary-based named entity recognition approach by checking if the corresponding unambiguous names appear within the same document. Nonetheless, they do not refer to this issue with regard to entity linking or in the case of missing contextual information on the sentence-level.

[1] We reported this unexpected behavior to the technical support team of the *Catalogue of Life* and have been notified that the sorting order of the results has been adapted to retrieve exact matches of Latin names. Regarding the vernacular names in the database, for the time being, no essential fixes can be done. (Personal correspondence with Wilfred Gerritsen from the *Catalogue of Life* technical support team, 20^{th} of February 2019.)

5.3 Web-Interface: End-to-End Named Entity Recognition and Linking

As a final step, we integrated the end-to-end pipeline including linguistic preprocessing, named entity recognition (NER) and entity linking (EL) into a web-interface.[2] We used Bootstrap for the implementation of the interface and Flask [Grinberg, 2014] to set up the web-application. Our main goal was to provide a simple interface that enables the user to input a raw text snippet in either German or English. After successful processing and tagging of the text, a JSON-object containing all the identified entity candidates is displayed and can be downloaded, if desired. If possible, all entity candidates proposed by the neural tagger are disambiguated using the customized lookup table for vernacular names and the *Catalogue of Life* (CoL) database for entity linking. The final JSON-file contains all the available taxonomic information associated with each unique entity.

For the detection, disambiguation, and interlinking of the candidate entities, we applied an end-to-end approach that can be sketched as follows:

(1) Tokenization of input text

(2) Re-merging of sentences split at botanical abbreviations

(3) Tagging of tokenized sentences using an adapted version of the script `tagger.py` [Lample et al., 2016] and the best-performing models for German and English

(4) Collection of positional information (sentence IDs and indices) for the candidate entities proposed by the LSTM-CRF tagger

(5) Run API queries for unique candidates (script `entity_linker.py`)

(6) Run lookups for vernacular names if the returned API response is empty, and, if the lookup is successful, re-run the API query for the disambiguated name

(7) Store the gathered information for each entity in a JSON-object

[2] Appendix D gives an overview on the scripts included in the web-interface processing pipeline for this end-to-end named entity recognition and linking approach. A publicly accessible development version runs on `https://imeran.pythonanywhere.com/`.

First, we prompt the user the choose the input language. The text snippet entered by the user is then segmented into sentences and tokenized (step (1) and (2)) and tagged in the background (step (3)) using the language-specific tokenizers provided by the *NLTK* [Loper and Bird, 2002] or *spaCy* [Honnibal and Montani, 2017] libraries. During tokenization, we additionally check for the presence of botanical abbreviations such as *var.* or *ssp.* and, if necessary, re-merge those sentences that have been split by mistake (step (2)). This interim step ensures that extensive infraspecies names such as "Lactuca sativa var. crispa" are not split and thus missed by the neural tagger. The adapted tagging script `tagger.py` [Lample et al., 2016] uses the best-performing language-specific *Wiki* model[3] to tag the tokenized text and produces annotations in IOB format as an output. In step (4), we iterate over each proposed entity candidate and collect positional information, such as the sentence ID and the index of the entity within a sentence. Finally, we run the API query for the unique entities using the *CoL* webservice [Roskov et al., 2018] in step (5). If the API response is empty, we check the language-specific lookup table and, if possible, disambiguate the concerned names using a Latin name in order to rerun the query. Especially for German vernacular names, this supplementary name lookups are indispensable to improve the overall linking coverage (see Table 29). As a final result, we return a JSON-object (step (7)) containing all entity candidates proposed by the neural tagger together with the associated positional and botanical information. The latter includes a unique database ID, the taxon rank, a URL to the *CoL* page of the taxon, the official name status, an associated accepted scientific name and, if specified, a bibliographic reference. On the web-interface, we display a prettified version of the JSON-object with an indentation size of two spaces in an output window (see Figure 6). We additionally make the JSON-file available for download.

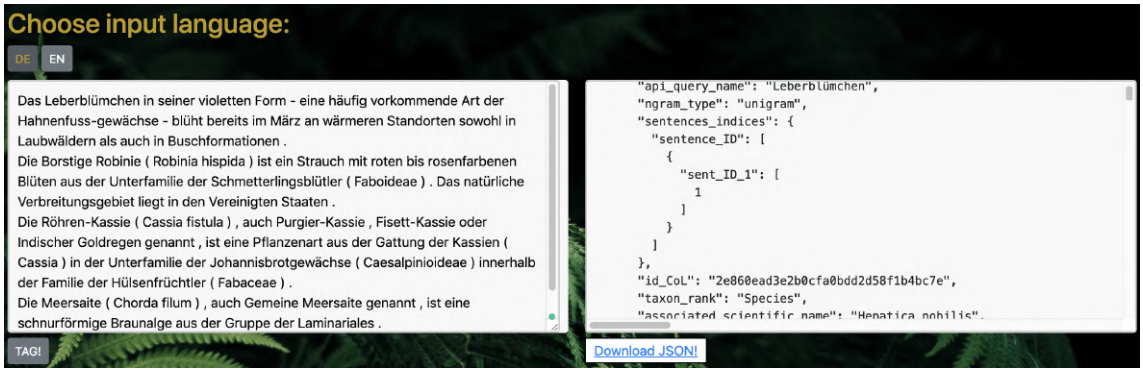

Figure 6: Project website for end-to-end named entity recognition and entity linking.

[3]For German, we used the *Wiki* model trained with a dropout rate of 0.3, whereas for English, we applied the model including the capitalization feature layer.

Admittedly, due to long compilation and model building times required by the script `tagger.py`, the web-interface is at this stage not optimized for fast performance. Equally, the API requests using the *CoL* webservice are time-intense. We prevent an excessive entity linking duration by querying only the unique entity candidates from the input text. This easy-to-use application is a demonstration for automatic entity extraction from raw, unstructured input text based on linguistic preprocessing, domain-specific rule-based corrections, named entity recognition and entity linking.

5.4 Summary and Discussion

Linking and disambiguating named entities is essential for information and knowledge extraction in manifold domains [Kolitsas et al., 2018]. Our entity linking evaluation on four distinct datasets showed that the integration of a tailored lookup table is valuable for the disambiguation and subsequent linking of German vernacular names. We achieved a maximum improvement in overall entity linking coverage of +16% for the German *PlantBlog* corpus containing a high concentration of vernacular names. For English, we reached a total linking coverage of maximally 70% of linked entities for the *TB* corpus using the *Catalogue of Life* (CoL) database and webservice. Despite being the most comprehensive and internationally accepted index of the world's known species, the disambiguation of non-English vernacular names is still hampered by incomplete databases and missing entries. The examples of false positive matches returned by the *CoL* webservice showed that there is still room for improvements to enhance automatic entity linking and disambiguation for vernacular names. Our experiments did, however, contribute to fixing the sorting order of the results on the level of scientific names in the CoL webservice. We hope that these insights encourage further work on international, taxonomic knowledge bases regarding the integration of vernacular botanical entities in lower-resourced languages.

6 Future Work and Outlook

Future tasks and challenges in this field involve the expansion of our presented approach on additional languages to recover and aggregate international biodiversity knowledge. Our experiments can serve as an exemplary pipeline for neural models trained on large-scale Wikipedia data and the subsequent application on lower-resourced text genres and new domains.

In terms of possible improvements, we suggest additional test runs using more coarse-grained entity labels. The evaluation presented by Nothman et al. [2013], for instance, shows a decrease of 4–6% in F_1-score when adopting fine-grained instead of coarse-grained entity types. Consequently, it could be beneficial to only distinguish between the categories "scientific name" and "vernacular name" in order not to confuse the neural tagger with the presence of numerous taxonomic labels. Yet, the detection of fine-grained entity labels on multiple hierarchical levels is particularly valuable for successfully classifying named entities and automatically populating or enriching taxonomic knowledge bases [Ekbal et al., 2010]. The inclusion of text material with condensed occurrences of low-frequent entity classes such as *phylum*, *class*, and *order* could additionally increase recall and the overall model performance on such taxonomic levels. Especially for lower-resourced domains and languages, the integration of context-sensitive embeddings [Akbik et al., 2018; Peters et al., 2018; Habibi et al., 2017] could additionally improve the performance and yield more accurate word representations for multilingual taxonomic entities.

With our work, we hope to contribute to existing biodiversity projects and standardization initiatives. The integration of botanical entities on a multilingual vernacular level into in-domain resources and databases is, in our opinion, highly important for safeguarding centuries-old plant knowledge. We want to emphasize that publicly available webservices providing a comprehensive query functionality for both scientific and vernacular name lookups can additionally contribute to revive the public interest in biodiversity and the world's botanical heritage, even for laypersons and not botanically-minded citizens. In this spirit, we hope to continue our research endeavors with the aim of bringing botany to the people through vernacular plant name knowledge.

7 Conclusion

In this thesis, we presented a semi-supervised approach for the recognition of scientific and vernacular plant names across different text genres for German and English. We created language-specific gazetteers to automatically annotate our training data using a dictionary-based tagging approach. Thus, we avoided time-consuming manual annotation without sacrificing high-quality training material. We found that adopting a fine-grained entity label set with a total of nine hierarchical labels for German, English and Latin results in robust performance across multiple genres. Our iterative approach allowed us to eliminate preprocessing errors as far upstream as possible in our annotation pipeline and to obtain higher quality data and a reduced performance gap between silver and gold standard evaluations. We compared eight bi-LSTM-CRF models [Lample et al., 2016] per dataset and language in single-dataset and cross-dataset evaluation settings. We found that low-effort models trained on large-scale Wikipedia abstracts can achieve a robust cross-corpus generalization ability and even outperform single-corpus models, which was the case for the German *PlantBlog* and the historical *BotLit* corpus. Generally, we observed the best performance on the Latin family label (`lat_fam`) obtaining F_1-scores of >99% for the German combined model and >98% for the English combined model. Regarding the performance on unseen entities from a related domain such as mycology, we want to highlight the beneficial effect of an additional capitalization feature layer in the bi-LSTM-CRF model [Lample et al., 2016]. After the second training round, we observed the best combined model performance on both silver-labeled and gold-labeled test sets with F_1-scores >94% for German and >86% for English. The best-performing German combined model uses a balanced dropout rate of 0.5, 300-dimensional pre-trained word embeddings, and a character embedding dimension of 29 for training. For the English combined dataset, adopting a dropout rate of 0.7 and a character embedding dimension of 25 outperforms the other parameter combinations. Concerning the entity linking evaluation, the integration of a customized, language-specific lookup table for disambiguating and subsequently linking vernacular names to a reference database resulted in a performance boost of maximally +16% in coverage for totally linked entities in the German datasets.

We hope that the generated data resources, large-scale *Wiki* models, and overall findings from this work will contribute to promote information extraction approaches in lower-resourced and under-explored domains, such as ethnobotany, folk medicine, and mycology. Our results underline the importance of fine-grained, hierarchical entity labels to enhance knowledge extraction in both scientific and non-scientific contexts. Our final project milestone for disambiguating and interlinking the entity candidates emphasizes the importance of integrating vernacular entities and associated knowledge into existing botanical knowledge bases. We hope that our approach can serve as a multilingual example for botanical end-to-end named entity recognition and encourage similar endeavors in other languages, in which precious traditional plant knowledge might be encoded.

Glossary

accuracy Common evaluation metric given by the ratio of correctly tagged tokens divided by the total number of tokens.

baseline An initial system used for comparison with subsequently trained and optimized systems.

corpus A (usually digitally available) collection of text material.

POS-tagging Part-Of-Speech (POS) tagging describes the process of automatically annotating word classes (part-of-speech, i.e. noun, adjective) in text.

gazetteer A domain-specific name list, sometimes also referred to as *dictionary*, used for name-lookups and dictionary-based annotation.

gold standard Manually labeled positive and negative examples in the data.

lemmatization A linguistic annotation step to find the *lemma*, that is, the canonical base form for a given token, i.e. the lemma of "houses" is "house".

machine learning Computational method to learn from previously seen input examples and to predict new, unseen instances.

named entity Any token or sequence of tokens constituting a name in the broader sense. This includes proper names, toponyms, taxonomic names etc.

neural network Learning architecture based on multiple layers with the goal to autonomously find patterns in the data and correctly predict new examples.

precision Common evaluation metric given by the ratio of correctly labeled tokens divided by the total number of labeled tokens.

recall Common evaluation metric given by the ratio of correctly labeled tokens divided by the total number of tokens that should have been labeled.

taxon A group of living species in taxonomy (plural *taxa*).

tokenization A linguistic preprocessing step involving the separation of single *tokens* at whitespaces and punctuation marking the end of a sentence.

References

Computational and Linguistic References

A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

O. Bender, F. J. Och, and H. Ney. Maximum Entropy Models for Named Entity Recognition. In *Proceedings of CoNLL-2003*, 2003.

C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked Data on the Web (LDOW2008). *WWW 2008 / Workshop Summary*, pages 1265–1266, April 2008.

N. Bubenhofer, M. Volk, F. Leuenberger, and D. Wüest. Text+Berg-Korpus (Release 151 Version 01), 2015.

R. Bunescu and M. Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 9–16, Trento, Italy, 2006.

X. Carreras, L. Màrquez, and L. Padró. Learning a perceptron-based named entity chunker via online recognition feedback. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, 4:156–159, 2003.

A. F. Chamberlain. The child and childhood in folk-thought. *Science*, 3:813–814, 1896.

G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan. An exploration of dropout with LSTMs. In *INTERSPEECH*, 2017.

C. Chiarcos, S. Hellmann, and S. Nordhoff. Towards a linguistic linked open data cloud : The open linguistics working group. *TAL*, 52(3):245–275, 2011.

N. Chinchor, E. Brown, L. Ferro, and P. Robinson. 1999 Named Entity Recognition Task Definition. *MITRE and SAIC*, 1999.

L. Chiticariu, Y. Li, and F. R. Reiss. Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832, Seattle, Washington, USA, 2013.

J. Cresswell. *Oxford Dictionary of Word Origins*. Oxford University Press, 2nd edition, 2010.

S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, Prague, 2007.

S. Cucerzan and D. Yarowsky. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *Proc. Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*, pages 90–99, 1999.

J. R. Curran, T. Murphy, and B. Scholz. Minimising semantic drift with Mutual Exclusion Bootstrapping. In *PACLING*, 2007.

M. Dredze, P. Mcnamee, D. Rao, A. Gerber, and T. Finin. Entity Disambiguation for Knowledge Base Population. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 277–285, Beijing, China, 2010.

A. Ekbal and S. Bandyopadhyay. Named Entity Recognition using Support Vector Machine: A Language Independent Approach. *World Academy of Science, Engineering and Technology*, 39:548–563, 2010.

A. Ekbal, E. Sourjikova, A. Frank, and S. P. Ponzetto. Assessing the challenge of fine-grained named entity recognition and classification. In *Proceedings of the 2010 Named Entities Workshop*, NEWS 2010, pages 93–101, Stroudsburg, PA, USA, 2010.

O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1):91–134, June 2005.

Getbootstrap.com. Bootstrap: Getting Started, 2015. URL `http://getbootstrap.com/getting-started/`. [Online; accessed 04-January-2019].

E. Grave. Weakly supervised named entity classification. In *Workshop on Automated Knowledge Base Construction (AKBC)*, Montréal, Canada, 2014.

E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

M. Grinberg. *Flask Web Development: Developing Web Applications with Python.* O'Reilly Media, Inc., 1st edition, 2014.

B. Hachey, W. Radford, and J. R. Curran. Graph-Based Named Entity Linking with Wikipedia. In A. Bouguettaya, M. Hauswirth, and L. Liu, editors, *Web Information System Engineering – WISE 2011. Lecture Notes in Computer Science*, pages 213–226. Springer, Berlin, Heidelberg, October 2011.

B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran. Evaluating entity linking with wikipedia. *Artificial Intelligence*, 194:130–150, January 2013.

J. Hammerton. Named Entity Recognition with Long Short-Term Memory. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, pages 172–175, 2003.

A. Hill. *The works of the late Aaron Hill, esq; ...: consisting of letters on various subjects, and of original poems, moral and facetious. With an essay on the art of acting.* Printed for the benefit of the family, 2 edition, 1754.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, 2012. URL `http://arxiv.org/abs/1207.0580`.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

Z. Huang, W. Xu, and K. Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, 2015. URL `http://arxiv.org/abs/1508.01991`.

M. Jansche. Named entity extraction with conditional Markov models and classifiers. In *Proceedings of the 6th Conference on Natural Language Learning - COLING-02*, volume 20, pages 1–4, Morristown, NJ, USA, 2002.

D. Jurafsky and J. H. Martin. *Speech and Language Processing (3rd Edition Draft)*. Prentice Hall, 2018.

N. Kolitsas, O. Ganea, and T. Hofmann. End-to-End Neural Entity Linking. *CoRR*, pages 519–529, 2018. URL `http://arxiv.org/abs/1808.07699`.

G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *HLT-NAACL*, 2016.

J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.

W. Ling, T. Luís, L. Marujo, R. Fernandez, A. S. Amir, C. Dyer, A. W. Black, and I. Trancoso. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal, 2015.

E. Loper and S. Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP 2002, pages 63–70, Stroudsburg, PA, USA, 2002.

M. Lui and T. Baldwin. Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, ACL 2012, pages 25–30, Stroudsburg, PA, USA, 2012.

X. Ma and E. H. Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *CoRR*, 2016. URL `http://arxiv.org/abs/1603.01354`.

M. Majlis. Wikipedia-API. GitHub repository, 2017. URL `https://github.com/martin-majlis/Wikipedia-API/`.

P. McNamee and H. T. Dang. Overview of the TAC 2009 Knowledge Base Population Track. In *Proceedings of Test Analysis Conference 2009 (TAC 09)*, 2009.

R. Mihalcea and A. Csomai. Wikify! Linking Documents to Encyclopedic Knowledge. In *CIKM 2007*, Lisboa, Portugal, 2007.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, 2013. URL `https://arxiv.org/pdf/1310.4546.pdf`.

A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014.

S. Morwal, N. Jahan, and D. Chopra. Named Entity Recognition using Hidden Markov Model (HMM). *International Journal on Natural Language Computing (IJNLC)*, 1(4), 2012.

D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

R. Navigli and A. Moro. Multilingual Word Sense Disambiguation and Entity Linking. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 5–7, Dublin, Ireland, 2014.

J. Ni and R. Florian. Improving Multilingual Named Entity Recognition with Wikipedia Entity Type Mapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1275–1284, Austin, Texas, 2016.

J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. Curran. Learning Named Entity Recognition from Wikipedia. *Artificial Intelligence*, 194:151–175, 2013.

G. Paliouras, V. Karkaletsis, G. Petasis, and C. D. Spyropoulos. Learning Decision Trees for Named-Entity Recognition and Classification. In *ECAI Workshop on Machine Learning for Information Extraction*, 2000.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

J. Raiman and O. Raiman. Deeptype: Multilingual entity linking by neural type system evolution. *CoRR*, 2018. URL `http://arxiv.org/abs/1802.01021`.

L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. *CoRR*, 1995. URL `http://arxiv.org/abs/cmp-lg/9505040`.

D. Rao, P. McNamee, and M. Dredze. Entity Linking: Finding Extracted Entities in a Knowledge Base. *Multi-source, Multilingual Information Extraction and Summarization, Theory and Applications of Natural Language Processing*, pages 93–115, 2013.

L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL 2009*, page 147, Boulder, Colorado, 2009.

E. F. T. K. Sang and J. Veenstra. Representing Text Chunks. In *EACL 1999 Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179, July 1999.

M. Sato, H. Shindo, I. Yamada, and Y. Matsumoto. Segment-Level Neural Conditional Random Fields for Named Entity Recognition. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 97–102, Taipei, Taiwan, 2017.

H. Schmid. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop.*, Dublin, Ireland, 1995.

W. Shakespeare. *Hamlet*. Edited by George Richard Hibbard. Oxford University Press, 2008.

W. Shen, J. Wang, and J. Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, February 2015.

K. Shinzato, S. Sekine, N. Yoshinaga, and K. Torisawa. Constructing dictionaries for named entity recognition on specific domains from the web. 2006.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

B. Strauss, B. E. Toma, A. Ritter, M.-C. De Marneffe, and W. Xu. Results of the WNUT16 Named Entity Recognition Shared Task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, 2016.

E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL 2003, pages 142–147, Stroudsburg, PA, USA, 2003.

M. Volk, N. Bubenhofer, A. Althaus, M. Bangerter, L. Furrer, and B. Ruef. Challenges in building a multilingual alpine heritage corpus. *Seventh International Conference on Language Resources and Evaluation (LREC)*, 2010.

V. Yadav and S. Bethard. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, USA, 2018.

W. Zhang, J. Su, C. L. Tan, and W. T. Wang. Entity Linking Leveraging Automatically Generated Annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1290–1298, Beijing, China, 2010.

Z. Zheng, F. Li, M. Huang, and X. Zhu. Learning to Link Entities with Knowledge Base. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 483–491, Los Angeles, California, 2010.

J. Zhou, B.-C. Li, and G. Chen. Automatically building large-scale named entity recognition corpora from Chinese Wikipedia. *Frontiers of Information Technology & Electronic Engineering*, 16(11):940–956, November 2015.

Botanical References

A. Achmüller. *Wickel, Salben und Tinkturen*. Raetia, 2nd edition, 2016.

D. Aeschimann and C. Heitz. Synonymieindex der Schweizer Flora und der angrenzenden Gebiete. 2. Auflage., 2005.

L. M. Akella, C. N. Norton, and H. Miller. NetiNeti: Discovery of scientific names from text using machine learning methods. *BMC Bioinformatics*, 13:211, 2012.

B. G. Bareja. Caveat: Common Plant Names. *Cropsreview*, 2010. URL `https://www.cropsreview.com/common-plant-names.html`.

N. Barrett and J. Weber-Jahnke. Building a biomedical tokenizer using the token lattice design pattern and the adapted Viterbi algorithm. *BMC Bioinformatics*, 12(Suppl. 3):S1, June 2011.

M. Basaldella, L. Furrer, C. Tasso, and F. Rinaldi. Entity recognition in the biomedical domain using a hybrid approach. *7th International Symposium on Semantic Mining in Biomedicine*, pages 11–19, 2017.

H. H. Bosshard. *Mundartnamen von Bäumen und Sträuchern in der deutschsprachigen Schweiz und im Fürstentum Liechtenstein*. Bühler Druck AG, Zürich, 1978.

B. Boyle, N. Hopkins, Z. Lu, J. A. Raygoza Garay, D. Mozzherin, T. Rees, N. Matasci, M. L. Narro, W. H. Piel, S. J. Mckay, S. Lowry, C. Freeland, R. K. Peet, and B. J. Enquist. The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics*, 14(1):16, Jan. 2013.

M. Brandon, J.-M. Moncalvo, and S. A. Redhead. Agaricales (Version 09 May 2007), 2007. URL `http://tolweb.org/Agaricales/20551/2007.05.09`.

P. D. Cantino, H. D. Bryant, K. de Queiroz, M. J. Donoghue, T. Eriksson, D. M. Hillis, and M. S. Y. Lee. Species names in phylogenetic nomenclature. *Systematic Biology*, 48(4):790–807, 1999.

R. D. Cundall. The meaning of the latin/greek names of some larger fungi. *NWFG Newsletter*, 1998.

GBIF. The Global Biodiversity Information Facility (GBIF). What is GBIF?, 2018. URL `https://www.gbif.org/what-is-gbif`.

M. Gerner, G. Nenadic, and C. M. Bergman. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85, February 2010.

M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):37–48, Jul 2017.

C. Helmut. *Die deutschen Pflanzen- und Tiernamen. Deutung und sprachliche Ordnung.)*. Quelle & Meyer, 1986.

H. E. Heß, E. Landolt, and R. Hirzel. *Flora der Schweiz und angrenzender Gebiete (Band 1, 2 und 3). 2. Auflage.* Birkhäuser AG, Basel, 1976.

M. C. Hogan and D. R. Taub. Plant. In C. J. Cleveland, editor, *Encyclopedia of Earth*. Environmental Information Coalition, National Council for Science and the Environment, Washington, D.C., 2011.

W. Höhn-Ochsner. *Pflanzen in Zürcher Mundart und Volksleben (Zürcher Volksbotanik)*. H. Rohr, Zürich, 1986.

IPNI. The International Plant Names Index, 2012. URL `http://www.ipni.org`.

IUCN. The iucn red list of threatened species. version 2018-2., 2018. URL `http://www.iucnredlist.org`.

S. Knapp. What's in a name? *Nature*, 408:33, 2000.

D. Koning, I. N. Sarkar, and T. Moritz. TaxonGrab: Extracting Taxonomic Names From Text. *Biodiversity Informatics*, 2:79–82, November 2005.

E. Landolt. *Unsere Alpenflora*. Schweizer Alpen-Club (SAC), 7th edition, 2003.

E. Landolt. *Flora des Sihltals von der Stadt Zürich bis zum Höhronen*. Fachstelle Naturschutz Kanton Zürich, Zürich, 2013.

P. Leary. Taxonfinder. GitHub repository, 2014. URL `https://github.com/pleary/node-taxonfinder`.

Lishuang Li, Liuke Jin, Zhenchao Jiang, Dingxin Song, and Degen Huang. Biomedical named entity recognition based on extended Recurrent Neural Networks. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 649–652. IEEE, November 2015.

I. Lobato Vila. Where do names of species come from? *All you need is biology*, September 2017.

G. Madaus. *Lehrbuch der biologischen Heilmittel*. Georg Thieme Verlag, Leipzig, 1938.

S. J. Meades. Definitions and Abbreviations of Terms used in the NOPD Checklist, 2018. URL `http://www.northernontarioflora.ca`.

A. Miles, N. Rogers, and D. Beckett. SKOS Core 1.0 Guide. World Wide Web Consortium, 2004. URL `http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/`.

Y. Minami, H. Takeda, F. Kato, I. Ohmukai, N. Arai, U. Jinbo, M. Ito, S. Kobayashi, and S. Kawamoto. Towards a Data Hub for Biodiversity with LOD. In H. Takeda, Y. Qu, R. Mizoguchi, and Y. Kitamura, editors, *Semantic Technology. JIST 2012. Lecture Notes in Computer Science*, pages 356–361. Springer, Berlin, Heidelberg, 2013.

A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-h. Liu, R. Torres, M. Krauthammer, W. W. Lau, H. Liu, C.-N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman. Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2):S3, September 2008.

D. Y. Mozzherin, A. A. Myltsev, and D. J. Patterson. "gnparser": a powerful parser for scientific names based on parsing expression grammar. *BMC Bioinformatics*, 18(1):279, May 2017.

N. Naderi, T. Kappler, C. J. O. Baker, and R. Witte. OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19):2721–2729, 2011.

M. Nebel and G. Philippi. *Die Moose Baden-Württembergs, Bd.1, Allgemeiner Teil (Grundlagenwerke)*. Ulmer, 2000.

C. Norton, I. N. Sarkar, and P. Leary. uBio Project, 2018. URL `http://www.ubio.org/`.

E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLOS ONE*, 8(6), 2013.

A. Pahler and K. Rücker. Schreibweise deutscher Pflanzennamen. *Gartenpraxis*, 27 (12):39–42, 2001.

D. J. Patterson, D. Remsen, W. A. Marino, and C. Norton. Taxonomic indexing: Extending the role of taxonomy. *Systematic Biology*, 55(3):367–373, 2006.

D. J. Patterson, J. Cooper, P. M. Kirk, and D. P. Remsen. Names are key to the big new biology. *Trends in ecology & evolution*, 25(12):686–691, 2010.

A. Pfeiffer. Einige oberösterreichische Trivialnamen der Pflanzen. (Vorgelegt in der Versammlung am 6. December 1898.), 1898.

C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser. AliBaba: PubMed as a graph. *Bioinformatics*, 22(19):2444–2445, 2006.

M. H. Porcher. Searchable World Wide Web Multilingual Multiscript Plant Name Database (MMPND) (online resource), 1999. URL `http://www.plantnames.unimelb.edu.au/`.

D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24:296–298, 2008.

H. R. Reinhard, P. Gölz, R. Peter, and H. Wildermuth. *Die Orchideen der Schweiz und angrenzender Gebiete*. Fotorotar AG, Egg, 1991.

M. Rickli. *Flora des Kantons Zürich*. Nägeli, Zürich, 2nd edition, 1912.

Y. Roskov, G. Ower, T. Orrell, D. Nicolson, N. Bailly, P. Kirk, T. Bourgoin, E. DeWalt, W. Decock, A. De Wever, E. Van Nieukerken, J. Zarucchi, and L. Penev. Species 2000 & ITIS Catalogue of Life, 24th September 2018, 2018. URL `www.catalogueoflife.org/col`.

G. Sautter, K. Böhm, and D. Agosti. A combining approach to find all taxon names (FAT). *Biodiversity Informatics*, 3, 2006.

M. A. F. Seideh, H. Fehri, and K. Haddar. Recognition and extraction of latin names of plants for matching common plant named entities. In L. Barone, M. Monteleone, and M. Silberztein, editors, *Automatic Processing of Natural-Language Electronic Texts with NooJ*, pages 132–144. Springer International Publishing, 2016a.

M. A. F. Seideh, H. Fehri, and K. Haddar. Named entity recognition from arabic-french herbalism parallel corpora. In T. Okrut, Y. Hetsevich, M. Silberztein, and H. Stanislavenka, editors, *Automatic Processing of Natural-Language Electronic Texts with NooJ*, pages 191–201, Cham, 2016b. Springer International Publishing.

V. Sharma, W. Law, M. J. Balick, and I. N. Sarkar. Harnessing Biomedical Natural Language Processing Tools to Identify Medicinal Plant Knowledge from Historical Texts. *Annual AMIA Symposium proceedings*, pages 1537–1546, 2017.

E. Sheng, S. Miller, J. L. Ambite, and P. Natarajan. A Neural Named Entity Recognition Approach to Biological Entity Identification. In *Proceedings of the BioCreative III Workshop*, pages 24–27, 2017.

P. Spescha. *Beschreibung der Val Tujetsch (1806)*. Chronos Verlag, Zürich, 2009.

J. H. Wiersema. GRIN Taxonomy. US National Plant Germplasm System., 2018. URL `https://doi.org/10.15468/ao14pp`.

Wikipedia. Agaricales - Wikipedia, The Free Encyclopedia, 2018a. URL `https://en.wikipedia.org/wiki/Agaricales`. [Online; accessed 04-February-2019].

Wikipedia. Boletales - Wikipedia, The Free Encyclopedia, 2018b. URL `https://en.wikipedia.org/wiki/Boletales`. [Online; accessed 04-February-2019].

Wikipedia. Liste der Gefässpflanzen - Wikipedia, The Free Encyclopedia, 2018c. URL `https://de.wikipedia.org/w/index.php?title=Liste_der_Gef%C3%A4ssapflanzen_Deutschlands/A&oldid=180876185`. [Online; accessed 04-November-2018].

Curriculum Vitae

Personal Details

Name	Isabel Meraner
DOB	16th of September, 1992
Address	Hürststrasse 11, CH-8046 Zurich
Email	isabel.meraner@uzh.ch

Education

2017–2019 **M.A**, *University of Zurich*, Zurich.
Multilingual Text Analysis

2012–2016 **B.A.**, *Ludwig-Maximilians-Universität*, Munich.
Romance and Scandinavian Linguistics

Work Experience

September 2018–now **Student Assistant**, *University of Zurich*, Zurich.
GeoKokos-project:
Project assistance, neural toponym recognition

April 2017–now **Student Assistant**, *University of Zurich*, Zurich.
Tutor of Massive Open Online Course (MOOC)
"Natural Language Processing for Digital Humanities" on Coursera

2015–2018 **Freelance Editor**, *Langenscheidt Publishing House*, Munich.
Editorial activities, multilingual neologism projects, lexicographic projects

2013–2016 **Student Assistant**, *Ludwig-Maximilians Universität*, Munich.
Homepage support, student advisory service, research work

Other

- Student representative of the Multilingual Text Analysis (MLTA) Program @ University of Zurich
- Engagement for Students Across Borders (German for Refugees)

A Lists

Full list of botanical works[1] (including translations) used for the extraction of vernacular plant names:

- "Flora der Schweiz" ('Flora of Switzerland') [Heß et al., 1976]
- "Flora des Sihltals ('Flora of Sihl Valley') [Landolt, 2013]
- "Flora des Kantons Zürich" ('Flora of the Canton of Zurich') [Rickli, 1912]
- "Mundartnamen von Bäumen und Sträuchern in der deutschsprachigen Schweiz und im Fürstentum Liechtenstein" ('Vernacular names of trees and shrubs of German-speaking Switzerland and Liechtenstein') [Bosshard, 1978]
- "Lehrbuch der biologischen Heilmittel" ('The textbook of biological remedies') [Madaus, 1938]
- "Unsere Alpenflora" ('Our alpine flora') [Landolt, 2003]
- "Pflanzen in Zürcher Mundart und Volksleben" ('Vernacular plant names in Zurich dialect and culture') [Höhn-Ochsner, 1986]
- "Die Orchideen der Schweiz und angrenzender Gebiete" ('Orchids of Switzerland and neighboring areas') [Reinhard et al., 1991]
- "Einige oberösterreichische Trivialnamen der Pflanzen" ('Vernacular plant names of Upper Austria') [Pfeiffer, 1898]
- "Die Moose Baden-Württembergs" ('Mosses of Baden-Wuerttemberg') [Nebel and Philippi, 2000]

[1] Most of the botanical works used for the present project have been kindly provided by the *plazi* institute. Alternatively, we retrieved the available, digitized works from the Internet.

B Tables

	German			English		
	P	R	F	P	R	F
Combined						
baseline	**95.11**	91.45	93.24	88.28	82.35	85.20
pre_emb	94.88	94.96	94.92	**88.60**	84.61	86.55
dropout 0.2	94.82	94.80	94.80	88.17	**85.90**	87.02
dropout 0.3	**94.88**	94.89	94.88	88.58	84.51	86.49
dropout 0.7	94.42	95.18	94.80	88.54	85.60	**87.04**
char_dim 29	94.94	94.98	**94.96**	88.32	85.19	86.73
char_dim 50	94.58	95.28	94.93	87.76	85.18	86.44
capdim 1	94.20	**95.55**	94.87	88.58	84.70	86.59
Wiki						
baseline	**95.06**	91.96	93.47	88.42	83.83	86.06
pre_emb	94.75	95.17	94.96	88.12	86.61	87.35
dropout 0.2	94.59	95.37	94.98	87.68	86.81	87.24
dropout 0.3	94.96	95.11	**95.04**	88.32	85.43	86.84
dropout 0.7	94.51	95.04	94.77	88.39	86.51	87.44
char_dim 29	94.60	95.31	94.95	**88.65**	**86.96**	**87.79**
char_dim 50	94.65	95.32	94.98	88.08	86.34	87.19
capdim 1	94.24	**95.71**	94.96	88.22	85.12	86.63
TB						
baseline	90.40	81.44	85.64	62.22	45.95	52.38
pre_emb	89.77	89.56	89.65	63.64	55.26	58.20
dropout 0.2	88.36	89.08	88.71	63.76	54.44	58.62
dropout 0.3	89.57	89.14	89.34	66.60	54.73	60.00
dropout 0.7	**90.55**	89.11	89.82	68.31	**56.23**	**61.31**
char_dim 29	90.04	89.31	89.65	66.47	53.42	59.18
char_dim 50	89.75	89.75	**89.75**	63.03	54.61	58.31
capdim 1	88.97	**89.88**	89.40	**71.06**	54.02	61.12
PlantBlog						
baseline	74.96	68.77	70.26	78.96	60.97	68.58
pre_emb	87.08	81.66	84.23	83.16	75.26	**78.99**
dropout 0.2	85.48	82.76	83.88	82.37	75.63	78.82
dropout 0.3	85.93	**83.36**	84.53	81.82	74.87	78.17
dropout 0.7	87.13	78.99	82.82	**86.16**	71.56	78.05
char_dim 29	**87.57**	80.85	83.96	84.90	73.43	78.71
char_dim 50	87.51	81.09	84.04	83.94	**76.97**	80.22
capdim 1	87.43	83.04	**85.07**	83.04	74.42	78.44
BotLit/S800						
baseline	0.00	0.00	0.00	89.04	80.89	84.69
pre_emb	60.09	21.01	28.19	89.59	84.61	86.99
dropout 0.2	61.52	35.94	44.43	89.21	83.30	86.13
dropout 0.3	67.81	**36.62**	**47.21**	90.14	82.98	86.39
dropout 0.7	40.00	6.50	11.11	**90.83**	**84.64**	**87.60**
char_dim 29	**71.08**	29.48	41.01	90.43	83.93	87.04
char_dim 50	55.00	14.96	22.44	89.04	83.43	86.10
capdim 1	66.59	27.82	38.24	89.84	83.87	86.71

Table 31: Evaluation results for all datasets (combined dataset, Wikipedia articles, mountaineering reports, blog articles, botanical literature) (see Section 3.1.3) and parameter combinations (see Section 4.2).

C Scripts

This appendix provides an overview on the Python scripts used for the single stages of this project. All scripts and data resources are also available from the GitHub repository `https://github.com/IsabelMeraner/BotanicalNER/`.

DATA COLLECTION (path = 'scripts/data_collection/')

```
# create Text+Berg subset of sentences containing plant names:
$ python3 get_subset_textberg.py -i ./../TextBerg/SAC/
  -o ./subset_textberg_de.txt
  -g ./../resources/gazetteers/ -l de

# generate Latin plant name abbreviations:
$ python3 add_latin_abbreviations.py -i
  ./../resources/gazetteers/lat/lat_species.txt
  -o ./outfile.txt

# generate German morpholocical variants:
$ python3 add_german_variants.py
  -i ./../resources/gazetteers/de/de_fam.txt
  -o ./outfile.txt

# split German compounds and add name variants:
$ python3 add_compound_variants.py
  -i ./../resources/gazetteers/de/de_species.txt
  -o ./outfileGAZ.txt

# create language-specific gazetteers:
$ python3 create_gazetteers.py
  -i ./../resources/gazetteers/de/de_species.txt -o outfile.txt

# add name variants to lookup-table:
$ python3 add_variants_database.py
  -i ./../resources/gazetteers/lookup_table/de_lat_referencedatabase.tsv
```

```
  -o ./outfile

# create fungi testset from Wikipedia articles:
$ python3 get_wiki_fungi_testset.py
  -o ./outfile.txt -c Pilze -l de

# retrieve Wikipedia abstracts and trivial names sections:
$ python3 retrieve_wiki_sections.py
  -i ./../resources/gazetteers/lat/lat_species.txt
  -t ./outfile_trivialsections.txt -a outfile_wikiabstracts.txt -l de

# extract plant names from Catalogue of Life archive:
$ python3 extracttaxa_cat_of_life -t ./colarchive/taxa/
  -v ./colarchive/vernacular/ -l ./latin.out
  -d ./german.out -e ./english.out -r rest_vernacular.out
```

PREPROCESSING (path = 'scripts/preprocessing/')

```
# tokenization:
$ python3 tokenize_corpus.py -d ./raw_data/ -l de

# part-of-speech tagging:
$ python3 ./treetagger-python_miotto/pos_tag_corpus.py
  -d ./../resources/corpora/
```

DICTIONARY-BASED ANNOTATION (path = 'scripts/annotation/')

```
# German annotation in IOB-format:
$ python3 iobannotate_corpus_de.py -d ./../resources/corpora/
  training_corpora/de/ -v ./../resources/gazetteers/de/
  -s ./../resources/gazetteers/lat/ -l de

# English annotation in IOB-format:
$ python3 iobannotate_corpus_en.py -d ./../resources/corpora/
  training_corpora/en/ -v ./../resources/gazetteers/en/
  -s ./../resources/gazetteers/lat/ -l de
```

TRAINING (path = 'scripts/training/')

```
# K-fold splitting of training data:
$ python3 kfold_crossvalidation.py
  -d ./../resources/corpora/training_corpora/de/

# Bashscript 5-fold crossvalidation training (examples):
$ bash bashscript_5foldtraining_preemb_en.sh
$ bash bashscript_5foldtraining_preemb_de.sh

# Adapted scripts from Lample et al.  (2016):
$ python2 train_no_dev.py
$ python2 utils.py
```

EVALUATION (path = 'scripts/evaluation/')

```
# Averaged evaluation over 5 folds:
$ python2 final_eval_kfold.py
  -d ./../../evaluation/baseline/model_baseline/
  -o ./evaluation_files/

# Evaluation of silver standard:
$ python3 evaluate_gold_silver.py
  -s ./../resources/corpora/gold_standard/
  de/alldata.test.fold1SILVER_de.txt
  -g ./../resources/corpora/gold_standard
  /de/alldata.test.fold1GOLD_de.txt

# Cross-dataset evaluation:
$ python3 cross_dataset_evaluation.py
  -s ./silver_standard/plantblog_corpus.test.fold1.txt
  -t ./tagged_data/model_wiki_test_blog_f1_dropout5.tsv

# File statistics training corpora (size, token, types, averaged length):
$ python3 file_statistics.py
  -i ./../resources/corpora/training_corpora/de/

# Transform IOB-format to 1-sentence-per-line (input for tagger.py):
$ python3 transform_iob_to_sentences.py
  -i ./../resources/corpora/training_corpora/
  de/botlit_corpus_de.tok.pos.iob.txt
  -o botlit_sentences.txt
```

ENTITY LINKING (path = 'scripts/entity_linking/')

```
# Catalogue of Life entity linking and creation of JSON-output:
$ python3 entity_linker.py
  -i ./../resources/corpora/training_corpora/de/
  botlit_corpus_de.tok.pos.iob.txt
  -o ./json_file.json -f IOB
  -r ./../resources/gazetteers/lookup_table/
  de_lat_referencedatabase.tsv -l True
```

D Web-Interface Pipeline

For running the web-application, the following dependencies need to be installed:

- Python3
- NLTK [Loper and Bird, 2002]
- spaCy [Honnibal and Montani, 2017]
- Flask for Python [Grinberg, 2014]

The following scripts and functions are included in the processing pipeline of the web-interface as described in Section 5.3.

```
# Start web-application:
$ python3 web_application.py

# Domain-adapted tokenization (function):
tokenize_input(inputText, language)

# Tagging of tokenized input sentence:
subprocess.call("python3 ./tagger-master/tagger.py -m ./models/{}
  -i ./output/input_tokenized.txt -o ./output/output_tagged.txt
  -d __".format(model), shell=True)

# Linking of entity candidates:
subprocess.call("python3 ./entity_linker.py
  -i ./output/output_tagged.txt -o ./static/output_linked.json
  --language {}".format(language), shell=True)
```

E Resources and Example Output

TRAINING DATA (path = 'resources/corpora/training_corpora/')

 # Silver standard training corpora (in IOB-format):

- plantblog_corpus_{de|en}.tok.pos.iob.txt
- wiki_abstractcorpus_{de|en}.tok.pos.iob.txt
- TextBerg_subcorpus_{de|en}.tok.pos.iob.txt
- {botlit|s800}_corpus_{de|en}.tok.pos.iob.txt

 # Gold standard fold of combined dataset (in IOB-format):

- combined.test.fold1GOLD_{de|en}.txt

 # Fungi testset for in-domain evaluation on held-out entities:

- test_fungi_{de|en}.tok.pos.iobGOLD.txt

GAZETTEERS (path = 'resources/gazetteers/')

 # Vernacular names (German):

- de_fam.txt
- de_species.txt

 # Vernacular names (English):

- en_fam.txt
- en_species.txt

 # Scientific names (Latin):

- lat_fam.txt
- lat_species.txt

- lat_genus.txt
- lat_subfam.txt
- lat_class.txt
- lat_order.txt
- lat_phylum.txt

 # Lookup tables for vernacular names:

- {de|en}_lat_referencedatabase.tsv

bi-LSTM-CRF MODELS (`path = 'resources/models/'`)

 # Best-performing models for German and English (single-dataset evaluation):

- model_combined_chardim29_de
- model_wiki_dropout0.3_de
- model_tb_dropout0.7_de
- model_plantblog_capdim1_de
- model_botlit_dropout0.3_de
- model_combined_dropout0.7_en
- model_wiki_chardim29_en
- model_tb_capdim1_en
- model_plantblog_chardim50_en
- model_s800_dropout0.7_en

 # Best-performing Wiki models for German and English (cross-dataset evaluation):

- model_wiki_crosscorpus_de_dropout0.3 (cross-corpus setting)
- model_wiki_crosscorpus_de_capdim1 (fungi test set)
- model_wiki_crosscorpus_en_preemb_dropout0.5 (cross-corpus setting)
- model_wiki_crosscorpus_en_capdim1 (fungi test set)

TAGGED DATA (path = 'resources/sample_output/')

 # Single-dataset model predictions:

- predictions_wiki_{de|en}.output
- predictions_tb_{de|en}.output
- predictions_plantblog_{de|en}.output
- predictions_{botlit|s800}_{de|en}.output

 # Cross-dataset model predictions:

- predictions_model_wiki_test_tb
  _preemb_chardim25_dropout5_capdim1_{de|en}.tsv
- predictions_model_wiki_test_plantblog
  _preemb_chardim25_dropout5_capdim1_{de|en}.tsv
- predictions_model_wiki_test_{botlit|s800}
  _preemb_chardim25_dropout5_capdim1_{de|en}.tsv

ENTITY LINKING (path = 'resources/linked_data/')

 # Vernacular-scientific lookup-table:

- {de|en}_lat_referencedatabase.tsv

 # Example JSON-output per data resource:

- json_data_wiki_{de|en}.json
- json_data_tb_{de|en}.json
- json_data_plantblog_{de|en}.json
- json_data_{botlit|s800}_{de|en}.json

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Philosophische Fakultät
Studiendekanat
Rämistrasse 69
CH-8001 Zürich
www.phil.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass die Masterarbeit von mir selbst ohne unerlaubte Beihilfe verfasst worden ist und ich die Grundsätze wissenschaftlicher Redlichkeit einhalte (vgl. dazu: http://www.uzh.ch/de/studies/teaching/plagiate.html).

Zürich, 01.03.2019 

.....

Ort und Datum Unterschrift