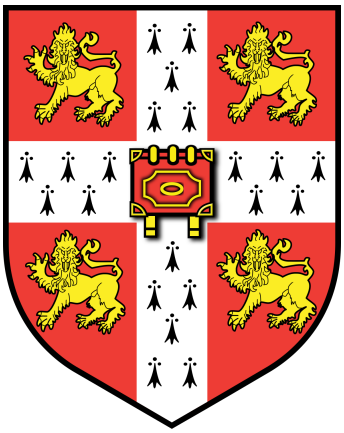
Die Entwicklung eines modernen Deutsch-Englisch Code-Switching-Datensatzes für Postings in sozialen Medien

Engineering Part IIA: SL3 Intermediate German

Igor Sterner (is473)

20. Juni 2022



Department of Engineering
Pembroke College, University of Cambridge

Betreuer: Alexander Bleistein
Constanze Leeb

Zusammenfassung

Dieser Bericht beschäftigt sich mit dem Problem, dass es bis jetzt keinen umfassenden Deutsch-Englisch Code-Switching (CSW)-Datensatz gibt. Um Sätze zu identifizieren, die CSW verwenden, braucht man eine Methode. Diese Methode ist nützlich bei transkribierten Gesprächen oder umgangssprachlichen Texten. Eine neue Methode ist in diesem Bericht implementiert. Ein Wörterbuch wird mit umgangssprachlichen Wörtern von dem „Urban Dictionary“ entnommen und mit englischen und deutschen Wörterbüchern kombiniert, um ein CSW-Wörterbuch zu entwickeln. Eine Wort-für-Wort-Überprüfung von 5,8 Mio. deutschen Tweets aus den ersten vier Monaten dieses Jahres hat 162 Tsd. Tweets als CSW identifiziert. Außerdem wurde der entwickelte Datensatz auf eine Genauigkeit von 77% geschätzt, was eine Basis für weitere Forschungsarbeiten sein kann.

Inhaltsverzeichnis

1 Einleitung

Natural Language Processing (kurz NLP) wird für verschiedene Disziplinen als interessant erachtet. In diesem Bereich unter Personen, die mehr als eine Sprache beherrschen, kann man in informellem Kontext ein häufiges Phänomen beobachten: das sogenannte Code-Switching (CSW) (Choudhury et al., 2019). Eigentlich im Bereich der herkömmlichen Sprachforschung anzusiedeln, rücken nun computergestützte Ansätze in den Fokus der NLP-Forschung. Deren Ziel ist es, die Leistung automatischer Spracherkennung und Übersetzungs-Programme mit CSW-Datensatzen zu verbessern (Yang et al., 2020; Adda-Decker et al., 2008). Ein Satz kann in der dominanten Sprache, auch ‚Matrix Language' (ML) genannt, formuliert sein. Diese wird durch das Hauptverb oder die Wortreihenfolge bestimmt. In solch einem Satz können darüber hinaus Wörter aus anderen Sprachen enthalten sein, der sogenannten ‚Embedded Language' (EL).

CSW findet statt, wenn sich die ML eines Satzes ändert, es aber keine EL gibt (Myers-Scotton, 1997). Im Gegenteil dazu lässt sich Code-Mixing in jedem Satz feststellen, in dem die EL zu finden ist. In diesem Bericht wird CSW für beide Situationen benutzt.

Den Code für diese Ansätze findet man unter
`https://github.com/igorsterner/CSW-Twitter`.

2 Stand der Forschung

Wie in Abb. 1 zu sehen ist, gab es bis 2014 fast keine computerlinguistischen CSW-Forschungsarbeiten, obwohl die erste CSW-Publikation in *ACL Anthology* bereits im Jahr 1998 erschien. Da CSW-Konferenzen alle zwei Jahre stattfinden, steigt die Anzahl an Publikationen seit 2014 in zweijährigen Zyklen. Es lässt sich beobachten, dass die Kurve zwischen 2016 und 2021 alle zwei Jahre um 100 Prozent ansteigt, was einen großes Interesse an der Thematik demonstriert.

Obwohl es viele Bestrebungen gibt, die Linguistik von CSW im Bezug auf Deutsch und Englisch zu erforschen (Müller et al., 2015; Eppler, 2010), gibt es leider nur zwei Forschungsarbeiten in *ACL Anthology*, welche „German" auch enthalten. Keine der beiden Forschungsarbeiten behandeln spezifisch das deutsch-englische CSW.

Das deutsch-englische Sprachpaar wurde bereits viel geforscht, zum Beispiel die Entwicklung von zweisprachigen Textkorpora, die oft als Trainingsdaten maschineller Übersetzungs-Modelle genuzt werden. Tatsächlich rangiert das Sprachpaar regelmäßig unter den fünf am meisten genutzten. Es ist daher erstaunlich, wie wenig Forschungsergebnisse bisher bezüglich CSW im deutsch-englischen Kontext existieren.

Es ist also unabdingbar, mehr im Bereich des deutsch-englischen CSW zu forschen, und ein moderner CSW Datensatz kann eine wichtige Grundlage für diese Arbeit bieten.

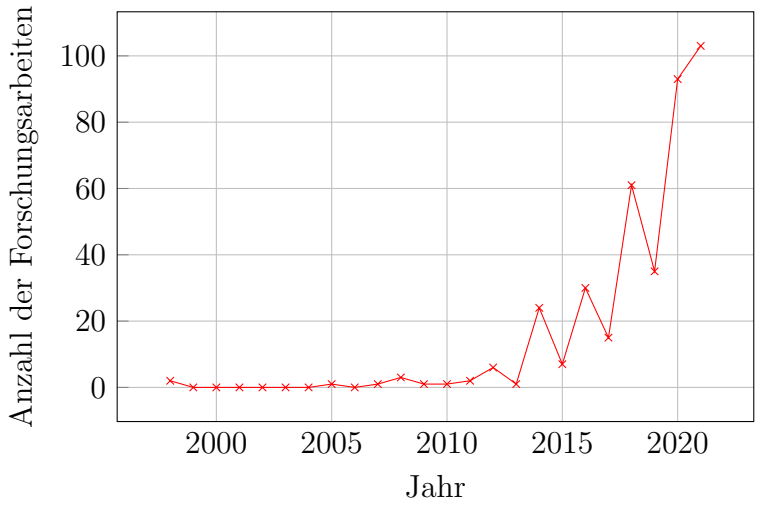


Abbildung 1: Forschungsarbeiten von der Zeitschrift *ACL Anthology*, die ‚Code-Switching', ‚Code-mixing' oder morphologische Varianten dieser Wörter im Titel oder in der Zusammenfassung enthalten

3 Code-Switching in den sozialen Medien

Traditionell wird CSW als mündliches Phänomen betrachtet. Die Suche nach Textkorpora, die regelmäßig für CSW verwendet werden, zeigt jedoch, dass informelle Chats aus den sozialen Medien eine ideale Quelle sind (Androutsopoulos and Hinnenkamp, 2001).

Die Nutzung von Postings in sozialen Medien hat zwei Hauptvorteile. Open-Source-Postings sind häufig genug, sodass es relativ einfach ist, einen umfangreichen Datensatz zu sammeln. Außerdem gibt es bereits viele NLP-Werkzeuge, um Tweets zu verarbeiten, von denen mehrere in diesem Bericht verwendet werden.

3.1 Datensätze für deutsche Postings in sozialen Medien

Viele Social-Media-Plattformen bieten Postingdaten zu Forschungszwecken an. Beispielsweise ist es möglich, Facebook-Posts (Franko, 2019), aber andere Datenquellen wie Reddit und Twitter können auch als Quelle für umgangssprachliche Sätze genutzt werden. Choudhury et al. (2019) zufolge wurde die Prävalenz der CSW in Tweets (Postings auf Twitter) auf 2-20% geschätzt . Nach der Freigabe der Twitter-API wurden viele große Twitter-Datensätze im Internet veröffentlicht (*Twitter API Documentation*, o. J.).

Ein Datensatz mit deutschen Twitter-Tweets, die über mehr als zwei Jahre mit der Twitter-API gesammelt wurden, ist online verfügbar (Kratzke, 2022). Der erste Schritt diese Projekts ist es, diese zu entpacken, zu extrahieren und zu bereinigen.

3.2 Extraktion der Twitter-Tweets

Der genutzte Datensatz beinhaltet Tweets, die in mehreren hunderten JSON-Dateien gespeichert werden. Diese wurden zunächst heruntergeladen. Ein wesentlicher Teil der Informationen war für dieses Projekt nicht relevant, zum Beispiel Autoreninformationen und Re-Tweet-Muster. Daher wurde im ersten Schritt der Pipeline der Text der Tweets extrahiert. Abb. 2 im Anhang zeigt das extraction.py-Skript, das diese Funktion ausführt.

Dass viele Tweets kein „Text"-Feld enthielten, sondern nur andere Metadaten, hat zu Schwieirgkeiten geführt. Dies wurde nicht weiter verfolgt, die entsprechenden Postings jedoch algorithmisch ignoriert.

3.3 Vorverarbeitung der Tweets

Bei der manuellen, visuellen Inspektion der Tweets wurde deutlich, dass ein Großteil der Postings aus URLS, Emojis usw. besteht. Außerdem wird in vielen Tweets das @-Symbol verwendet, um andere Twitter-Benutzer zu erwähnen. Da diese Informationen für das Projekt nicht relevant waren, wurden alle Wörter, die @-Symbol enthalten, entfernt.

Im Vergleich zum Spracherkennungsalgorithmus von Twitter führte die Bereinigung der Tweets, vor der Spracherkennung, zu einer besseren Leistung bei der Entfernung englischer Tweets. Weil fast alle dieser Wörter im CSW-Wörterbuch enthalten waren, wie in Kapitel 4.1 erklärt ist, war die Qualität des ersten generieten Datensatzes schlecht. Also wurde eine zeitaufwändige, aber effektive Spracherkennungsmethode aus der langdetect-Bibliothek implementiert, um zu überprüfen, ob der Tweet tatsächlich deutsch ist. Die bessere Leistung dieser Implementierung gegenüber der von Twitter ist wahrscheinlich auf die Bereinigung der Tweets zurückzuführen, bevor die Spracherkennung durchgeführt wurde. Insgesamt wurden 23,8% der Tweets hier entfernt, weil sie nicht als deutsch identifiziert wurden.

In Abb. 2 im Anhang sind die beiden Effekte des preprocessing.py-Skripts dargestellt. Zuerst wurden @s, Emojis und URLs entfernt, wie in Tweet (2) gezeigt. Zweitens wurde der Tweet (3) nach der Textbereinigung als englischer Text identifiziert und daher entfernt.

4 Cie Erkennung von Code-Switching in Tweets

Um einen neuen Deutsch-Englisch-CSW-Erkennungsalgorithmus zu entwickeln, muss eine Quelle der relevanten englischen Wörter gefunden werden. Dann muss ein Wort-für-Wort-CSW-Überprüfungsalgorithmus implementiert werden.

Das Flussdiagramm in Abb. 3 im Anhang stellt den CSW-Erkennungsalgorithmus dar, der in diesem Projekt umgesetzt wird. Jedes Wort in jedem bereinigten Tweet wird mit

einem speziellen CSW-Wörterbuch verglichen und wenn der Tweet mindestens ein CSW-Wort enthält, wird er gespeichert.

4.1 Wörterbücher

Einer der komplexesten Aspekte dieses Projekts war die Entwicklung eines speziellen CSW-Wörterbuchs für das Sprachpaar ‚DE-EN'.

In diesem Bericht wird der folgende Ansatz zur Erstellung dieses Wörterbuchs vorgestellt: Man nimmt ein kleines Wörterbuch mit häufig verwendeten englischen Wörtern und entfernt alle Wörter, die auch in einem großen Wörterbuch mit deutschen Wörtern vorkommen. Tab. 1 gibt Aufschluss über die Anzahl der Wörter von Wörterbüchen, die in diesem Projekt verwendet wurden.

Wörterbuch	**Quelle**	**Anzahl der Wörter**
Englisch	WL 100K Sätze (2020)	15K
Urban Dictionary	Bierner (2022)	13K
Deutsch 1	dict.cc	583K
Deutsch 2	WL 300K Sätze (2021)	65K
Schweizer	WL 300K Sätze (2012)	37K
Österreichisch	WL 100K Sätze (2012)	34K

Wörterbuch	**Anzahl der Wörter**
Gesamt-Englisch	22K
Gesamt-Deutsch	749K
CSW	**12K**

Tabelle 1: Verwendete Wörterbücher mit entsprechender Größe

4.1.1 Englische und deutsche Wörterbücher

Der ‚Wortschatz Leipzig' (WL) bietet einsprachige Textkorpora aus Web-Nachrichten an, zusammen mit zugehörigen Wörterbüchern für alle darin enthaltenen Wörter (Goldhahn et al., 2012). Das Ergebnis ist ein modernes und effizientes Wörterbuch für die gängigsten Wörter, die in geschriebenen Texten verwendet werden.

Im Laufe des Projekts wurde erkannt, dass ein Großteil der Tweets in deutschen Dialekten geschrieben werden (siehe Kapitel 5.3). Der ‚Wortschatz Leipzig' war sehr nützlich, um dieses Problem zu beheben, weil für die schweizerischen und österreichischen Nachrichten spezialisierte Textkorpora mit entsprechenden Wörterbüchern zur Verfügung standen. Für jedes Wörterbuch war es möglich, wenig verwendete Wörter zu entfernen. Dies begrenzt die Menge an eingeführten englischen Wörtern. Es war erforderlich, unterschiedlich streng ‚wenig verwendet' für unterschiedliche Dialekte zu quantifizieren. Die ausgewählte minimale Verwendungsfrequenz ist für jeden Dialekt auf der Github-Seite enthalten.

Obwohl es aus WL-Wörterbüchen bis zu 1 Mio. Sätze gibt, reicht diese Anzahl nicht aus, um obsukure deutsche Wörter aus dem CSW-Wörterbuch zu entfernen. Ein sehr

großes deutsches Wörterbuch war erforderlich, um Wörter wie „Digga“ zu entfernen. Glücklicherweise kann man das gesamte dict.cc-Wörterbuch herunterladen, also wurde dieses mit einem mittelgroßen Wörterbuch von WL für deutsche Wörter verwendet.

4.1.2 ‚Urban Dictionary‘

Die Verwendung von abgekürzten englischen Wörtern wie „idk“ oder anderer englischer Umgangssprache ist in deutschen Postings in sozialen Medien üblich. Deshalb war ein spezielles Wörterbuch der englischen Umgangssprache erforderlich, um diese Wörter erfolgreich als Verwendungen von CSW zu identifizieren.

Um dieses Problem zu lösen, wurde das ‚Urban Dictionary‘ verwendet. Das moderne ‚Urban Dictionary‘ enthält umgangssprachliche Texte, die in traditionellen englischen Wörterbüchern nicht enthalten sind. Alle Wörter wurden aus ihren Phrasen getrennt. Großgeschriebene Wörter, Wörter mit weniger als drei Zeichen und Wörter, die nicht alphabetische Zeichen enthalten, wurden entfernt. Die Wörter die mehr als zehnmal im ‚Urban Dictionary‘ vorkommen, wurden gespeichert.

4.1.3 Code-Switching-Wörterbuch

Schlussendlich wurde die Menge aller deutschen Wörter (ohne Wiederholungen) aus der Menge der englischen und ‚Urban Dictionary‘-Wörter entfernt, um ein ‚CSW-Wörterbuch‘ zu erstellen.

Um das Ziel zu erreichen, war es beabsichtigt, ein so großes deutsches Gesamtwörterbuch wie möglich zu erstellen, ohne versehentlich viele englische Wörter zu integrieren. Gleichzeitig war ein kleines englisches Wörterbuch wünschenswert, um den Rechenaufwand für die Suche nach CSW-Wörtern in Tweets zu verringern. Die verwendete subtraktive Methode bedeutete, dass die Größe des CSW-Wörterbuchs nicht größer als die ‚Gesamt-Englisch‘-Wörter sein konnte. Es wurden 44% dieser Wörter entfernt, weil es ‚identische‘ deutsche Wörter oder Wörter in deutschen Dialekten gibt.

4.2 Wort-für-Wort-Überprüfung der Tweets

Der finale Schritt in Abb. 2 im Anhang zeigt die Ergebnisse des CSW-Erkennungsalgorithmus. Tweet (1) hat CSW und (2) ist einfach deutsch. Dieser Algorithmus identifizierte das Wort ‚swear‘, das sich aus der ersten CSW-Phrase ‚I swear to god‘ ergibt, und die englische Abkürzung ‚idk‘ (englisch: I don’t know) als Verwendungen von CSW.

Das Flussdiagramm in Abb. 3 in Anhang enthält zwei Datenquellen: die Wörterbücher und die JSON-Dateien der Tweets. Die CSW-Wörterbücher werden wie in Kapitel 4.1 beschrieben erstellt und die Tweets werden wie in Kapitel 3.3 vorgestellt bereinigt.

Wort	Frequenz
btw	1,66%
finds	1,12%
wtf	1,02%
did	0,83%
idk	0,80%
understand	0,77%
been	0,74%
cringe	0,65%
weird	0,65%
dont	0,64%

Tabelle 2: Die 10 CSW-Wörter, die am häufigsten in Tweets verwendeten werden

Zunächst wird geprüft, ob das Wort in Großbuchstaben geschrieben ist, was auf Namen oder andere Eigennamen hinweist. Wenn das der Fall ist, soll es nicht als CSW-Wort identifiziert werden. Dann wird das Wort mit dem CSW-Wörterbuch abgeglichen, und wenn ein Tweet ein bestimmtes CSW-Wort enthält, wird er gespeichert.

5 Ergebnisse

Der vorgestellte Algorithmus wurde auf Tweets aus den ersten vier Monaten des Jahres 2022 angewendet. Von 5,8 Mio. bereinigten Tweets wurden 162 Tsd. als CSW-Tweets identifiziert. Dies entspricht einem Prozentsatz von 2,8. Im Kapitel 3.1 wurde bereits dargestellt, dass die Prävalenz der CSW in Tweets auf 2-20% geschätzt wurde. Obwohl unser Ergebnis am unteren Ende liegt, liegt es trotzdem innerhalb der erwarteten Prozentsätze.

Tab. 2 zeigt die 10 Wörter, die in diesen Tweets am häufigsten als Verwendung von CSW identifiziert wurden. Die ersten, dritten und fünften Abkürzungen, „btw“, „wtf“und „idk“, sowie die achthäufigste, „cringe“, sind nicht im englischen Originalwörterbuch enthalten. Dies unterstreicht die Wichtigkeit der neuartigen UD Benutzung, die in diesem Bericht vorgestellt wird.

5.1 Ein Open-Source-Datensatz

Dieses Projekt hatte zum Ziel, eine Methode zur Erstellung eines großen Datensatzes von deutsch-englischen CSW-Sätzen zu entwickeln.

Die allgemeinen Geschäftsbedingungen von Twitter verhindern das öffentliche Hochladen des gesamten Datensatzes, aber alle erforderlichen Quellen für diese Methode sind in der README.md von Github übersichtlich aufgelistet. Deshalb ist der Datensatz leicht reproduzierbar.

5.2 Menschliche Evaluierung

Eine Doktorandin, die fließend Deutsch und Englisch spricht, erhielt je 100 Tweets aus dem bereinigten Originaldatensatz und 100 Tweets, in denen CSW identifizert wurden. Sie wurde gebeten, alle Tweets zu identifizieren, in denen CSW verwendet wurde. 'Hashtags' mit englischen Wörtern wurden nicht als CSW identifiziert. Weil diese menschliche Evaluierung nur 100 Tweets verwendet und das eine kleine Probe von 5,8 Mio. Tweets ist, kann man nur erste Ergebnisse erreichen.

Von 100 zufällig ausgewählten Tweets wurden 13 als CSW-Tweets identifiziert. Dies entspricht einem Prozentsatz von 13. Das ist genau in der Mitte des erwarteten Bereichs (Choudhury et al., 2019). Da die Methode dieses Berichts nur 3% als CSW-Tweets identifiziert hat, ist es auch eine konservative Methode, CSW-Sätze zu identifizieren. Auf der anderen Seite wurden von 100 Tweets, die durch den Algorithmus als CSW-Tweets gekennzeichnet wurden, 77 korrekt als CSW-Tweets identifiziert. Dies entspricht einem Korrektheit-Prozentsatz von 77, was relativ gut ist.

5.3 Begrenzungen

Die Hauptbegrenzung dieses Berichtes ist die Behandlung von Wörtern wie ‚die': es könnte sich um den deutschen Artikel oder das englische Wort für ‚sterben' handeln. Solche Wörter werden in der vorhandenen Studie nicht betrachtet: Sie können nicht als CSW-Wörter identifiziert werden. In der praktischen Umsetzung führte dies dazu, dass 44% der Wörter des englischen UD entfernt wurden. Diese Studie unterschätzt daher die Verbreitung von CSW auf Twitter.

Besonders problematisch wares, englische Wörter zu vermeiden, die in den verschiedenen deutschen Dialekten gleich geschrieben werden. Leider bietet keine der untersuchten Spracherkennungsbibliotheken Funktionen zur Erkennung spezifischer deutsche Dialekte. Der Twitter-Spracherkennungsalgorithmus kategorisiert alle deutschen Dialekte lediglich mit ‚DE'. Um dieses Problem zu lösen, wurden spezialisierte schweizerische und österreichische Deutschwörterbücher eingeführt, da ein großer Teil der Tweets in diesen beiden Sprachen identifiziert wurde. Trotzdem enthält der finale Datensatz viele Tweets mit weniger verbreiteten Dialekten, die falsch eine CSW-Bedingung auslösen. Zum Beispiel wurde Luxemburgisch oft als CSW-Englisch erkannt, oft für Wörter wie ‚deen', das im Urban Dictionary viele ungewöhnliche Definitionen hat.

Markennamen, Bücher und Filmnamen wurden bei der Bewertung nicht als Verwendungen von CSW betrachtet. Es wurde versucht, großgeschriebene Wörter nicht zu betrachten, wie in Abb. 3 im Anhang beim Schritt vor der Prüfung mit dem CSW-Wörterbuch zu sehen ist. Trotzdem ist eine falsche Großschreibung in Postings in sozialen Medien üblich, zum Beispiel nur das erste Wort großzuschreiben. Deshalb lösen diese fälschlicherweise den CSW-Erkennungsalgorithmus aus.

Eine weitere wichtige Begrenzung des derzeitigen Algorithmus ist die Fähigkeit des Algorithmus Rechtschreibfehler zu erkennen. Rechtschreibfehler *sine* aus zwei Gründen *prob-*

limatisch (zwei häufige Fehler). Erstens führen sie oft dazu, dass der Spracherkennungsalgorithmus falsch ‚DE' ausgibt, wenn es in Wirklichkeit Englisch ist. Zweitens wurden Rechtschreibfehler entdeckt, die dazu führen, dass andere Wörter wie CSW-Wörterbuch-Wörter geschrieben werden.

6 Fazit

In diesem Bericht wird eine Methode zur Entwicklung eines großen Datensatzes von CSW-Sätzen mit einer mittelhohen Korrektheit beschrieben. Unter Berücksichtigung der verfügbaren Zeit und Ressourcen ist dieser Datensatz ein nützliches Ergebnis. Der generierte Datensatz kann die Basis für weitere Forschungsarbeiten sein. Dieser Datensatz könnte auch für andere NLP-Aufgaben benutzt werden. Der Autor hofft, dass diese Arbeit zu weiterer Forschung in dem Felt von Deutsch-Englisch-CSW anregt.

Der vorgestellte Algorithmus ist leider auch mit Begrenzungen verbunden. Die falsche Identifizierung von englischen Tweets stellt immer noch ein großes Problem für die Konsistenz dieses Datensatzes dar. Insbesondere zum Training Data, ist es für NLP-Modelle wichtig sicher zu sein, dass jeder Satz Deutsch Englisch CSW verwendet. Die innovative Verwendung des Urban Dictionary erlaubt die Erstellung eines modernen Deutsch-Englisch CSW-Datensatzes für Postings in sozialen Medien, zum Beispiel um moderne englische Abkürzungen zu erkennen.

Weitere Bereinigung des Datensatzes oder ein adaptiverer Erkennungsalgorithmus wird für verbesserte CSW-Genauigkeit in Zukunft notwendig sein. Bessere Bereinigung der Tweets würde mehr Tweets entfernen, die in verschiedenen Dialekten geschrieben sind. Außerdem könnte das CSW-Wörterbuch größer sein, um Wörter aus mehr Ländern aufzunehmen, in denen die Hauptsprache ein Dialekt des Deutschen ist. Mit einer statistischen Methode wäre es für den Algorithmus möglich, den Kontext besser zu erkennen.

7 Literaturverzeichnis

Adda-Decker, M., Pellegrini, T., Bilinski, E. and Adda, G. (2008). Developments of "Lëtzebuergesch" Resources for Automatic Speech Processing and Linguistic Studies, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco.
URL: *http://www.lrec-conf.org/proceedings/lrec2008/pdf/855_paper.pdf*

Androutsopoulos, J. and Hinnenkamp, V. (2001). Code-Switching in der bilingualen Chat-Kommunikation: ein explorativer Blick auf #hellas und #turks, *Chat-Kommunikation: Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation.* S. 67–401.

Bierner, M. (2022). Urban Dictionary List.
URL: *https://github.com/mattbierner/urban-dictionary-word-list*, Abruf am: 2022-05-23.

Choudhury, M., Srinivasan, A. and Dandapat, S. (2019). Processing and Understanding Mixed Language Data, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Association for Computational Linguistics, Hong Kong, China.

Eppler, E. (2010). *Emigranto: The syntax of German-English code-switching*, Wilhelm Braumüller Universitäts-Verlagsbuchhandlung.

Franko, K. (2019). Code-Switching in der computervermittelten Kommunikation. Eine Analyse deutsch-italienischer Facebook Beiträge., , Abruf am: 2022-05-23. *Ludwig-Maximilians-Universität München* .
URL: *http://www.kit.gwi.uni-muenchen.de/?p=4442&v=1*

Goldhahn, D., Eckart, T. and Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, S. 759–765.
URL: *http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf*

Kratzke, N. (2022). Monthly Samples of German Tweets. Type: dataset.
URL: *https://zenodo.org/record/6528564/export/hx*, Abruf am: 2022-05-23.

Myers-Scotton, C. (1997). *Duelling Languages: Grammatical Structure in Codeswitching*, Clarendon Press.
URL: *https://books.google.co.uk/books?id=NuYdnTyKkdQC*

Müller, N., Gil, L., Eichler, N., Geveler, J., Hager, M., Jansen, V., Patuto, M., Repetto, V. and Schmeißer, A. (2015). *Code-Switching: Spanisch, Italienisch, Französisch. Eine Einführung*, narr studienbücher, Narr Francke Attempto Verlag.
URL: *https://books.google.co.uk/books?id=P213DwAAQBAJ*

Twitter API Documentation (o. J.).
URL: *https://developer.twitter.com/en/docs/twitter-api*, Abruf am: 2022-06-06.

Yang, Z., Hu, B., Han, A., Huang, S. and Ju, Q. (2020). CSP:Code-Switching Pre-training for Neural Machine Translation, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, S. 2624–2636.
URL: *https://aclanthology.org/2020.emnlp-main.208*

A Anhang

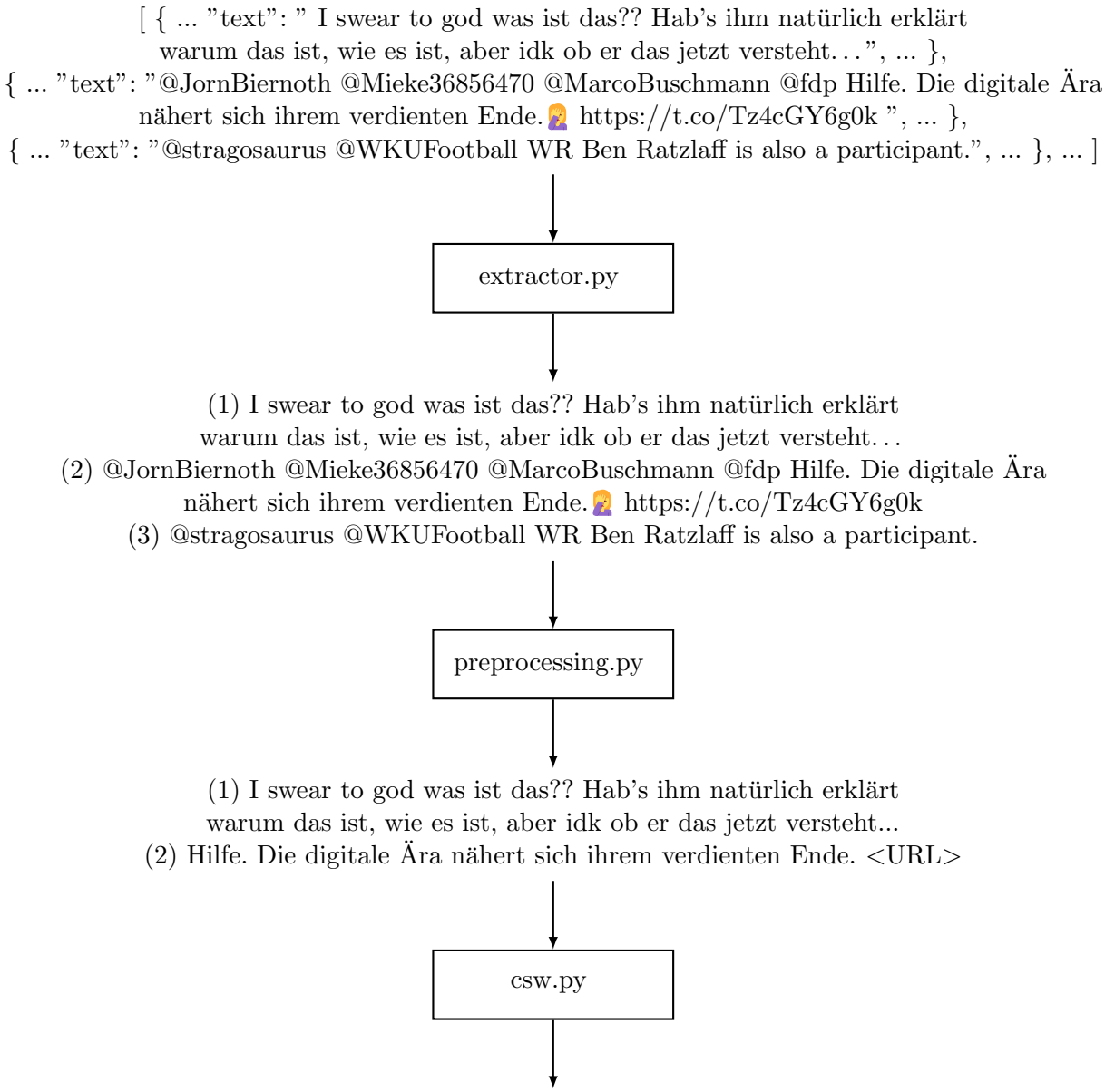


Abbildung 2: Ein Beispiel für den Extraktions-, Vorverarbeitungs- und Identifizierungsschritte

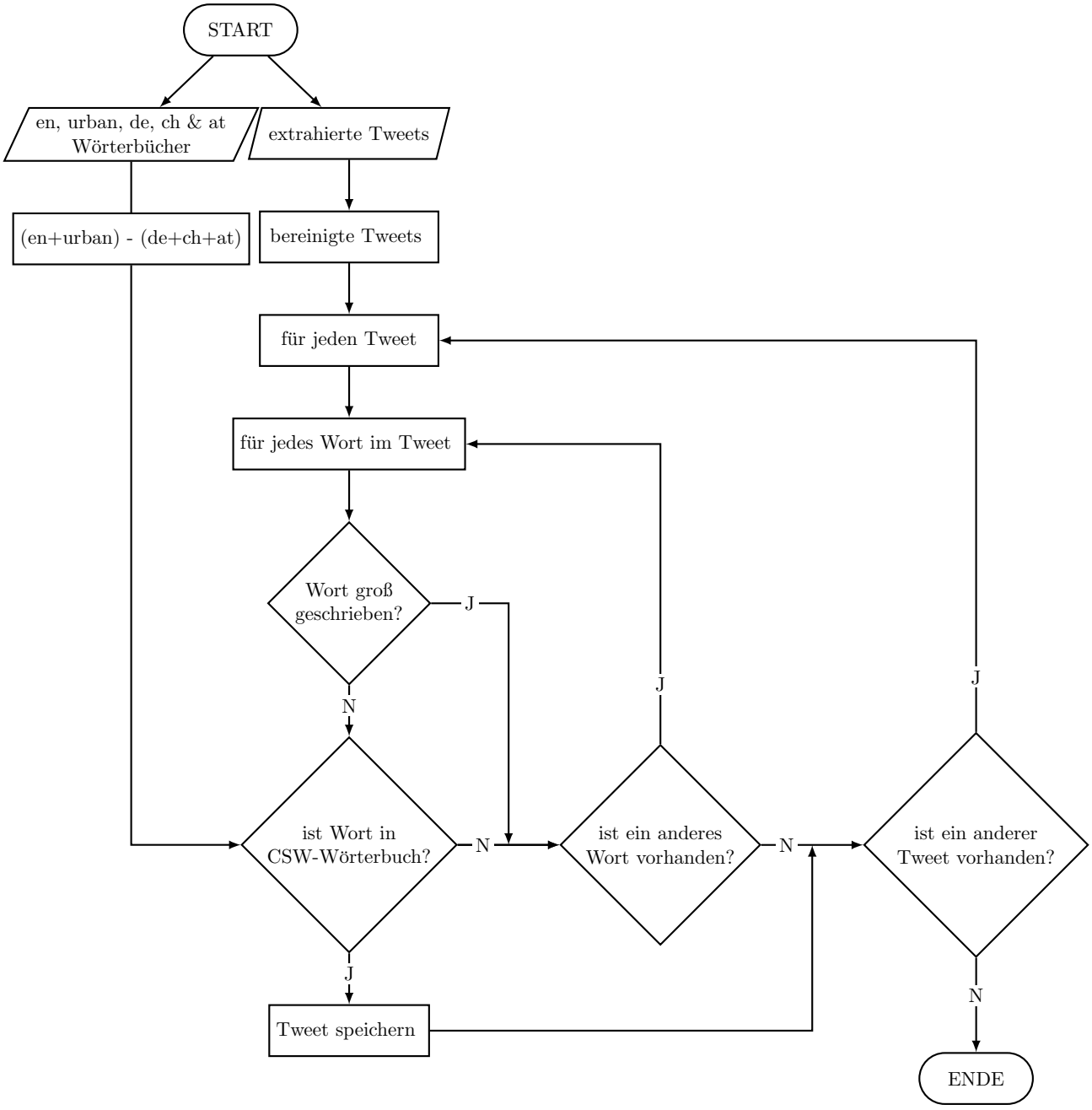


Abbildung 3: Ein Flussdiagramm zur Darstellung des CSW-Erkennungsalgorithmus