CHUN NR. 35/2020, S. 36–57

Zu Problematik und Verbesserungsmöglichkeiten der maschinellen Übersetzung von Wortbildungen des Deutschen und Chinesischen

WANG Kai 王锴

Zusammenfassung

Zwischen der deutschen und der chinesischen Sprache, die unter schriftlichen, grammatischen und kulturellen Aspekten weit voneinander entfernt sind und für deren maschinelle Übersetzung (MÜ) in die jeweils andere Sprache nicht genügend bilinguale Parallelkorpora existieren, treten falsche oder unverständliche Übersetzungen relativ häufig auf. Übersetzungsfehler bei unregistrierten Wörtern und Ad-hoc-Bildungen stellen einen Großteil davon dar.

Ziel der vorliegenden Arbeit ist herauszufinden, warum und wie künstliche Intelligenz (KI) bei der schriftlichen Übersetzung von Wortbildungen des Deutschen und des Chinesischen Fehler macht. Warum sind Segmentationen der Wörter nötig und wie laufen sie in den deutschen und chinesischen Korpora ab? Wie kann KI anhand von annotierten Korpora, bekannten Wörtern und sprachlichen Regeln qualitativ höherwertige Übersetzungen von unbekannten Wortbildungen zwischen den beiden Sprachen anbieten?

Keywords: DE-ZH-MÜ, maschinelle Übersetzung, Alignment, Wortbildungen, Segmentation

1 Einleitung

1.1 Hintergrund: Zur Bedeutung der maschinellen Übersetzung zwischen der deutschen und der chinesischen Sprache (DE-ZH-MÜ)

Die zunehmende Verwendung maschineller Übersetzung (MÜ) in diversen Branchen und gesellschaftlichen Kontexten erfordert immer mehr eine hochwertige Qualität solcher Übersetzungen. Die Fehler der MÜ können erhebliche Missverständnisse verursachen oder den Sinn gänzlich entstellen. Auch für fremdsprachendidaktische Bedürfnisse sind bessere Übersetzungen von Vorteil: Übersetzungsfehler können Fremdsprachler*innen bspw. durch automatische Übersetzer (Autoübersetzer) beim Selbstlernen des Chinesischen auf die falsche Fährte führen. Umgekehrt kann MÜ auch Deutschlernende beim Erwerb von Komposita verwirren, bspw. bei der Frage, ob es ein Fugenelement zwischen Wortgliedern gibt oder ob zwei Wörter zusammen-

schrieben werden: Google etwa zeigt die deutschen Komposita oft in nicht authentischer, sondern in phrasenähnlicher oder „verenglischter“ Form.

Die Problematik der DE-ZH-MÜ ist vielschichtig. In den heutzutage beliebtesten MÜ-Systemen werden häufig hybride statistik-basierte (SMT) und neuronale (NMT) Technologien angewendet, wie z. B. bei Google Translate, Bing Translate und Baidu Fanyi (vgl. Ge et al. 2018: 564). Wegen der stetig steigenden Datengrößen und für eine effizientere Übersetzung wird das Alignment[1] (die Zuordnung einander entsprechender Ausdrücke in zwei oder mehreren Sprachen) der multilingualen Parallelkorpora meistens *top-to-bottom* durchgeführt und immer mehr von den kontextabhängigen Algorithmen bedingt. Je größer die eingegebene sprachliche Einheit ist, desto besser kann die künstliche Intelligenz (KI) eine umso bessere Übersetzungsqualität erzielen. Umgekehrt gilt aber: Je kleiner eine sprachliche Einheit ist, desto höher kann die Fehlerdichte sein. Bei kontextfreien Wörtern, unter denen verschiedene einzelsprachliche und interlinguale Ambiguitäten existieren, treten deswegen am häufigsten Fehler auf. Im Prozess des Fremdsprachenlernens ist aber das Nachschlagen einzelner Wörter oder fester Phrasenkollokationen in der Regel viel gefragter als die Übersetzung ganzer Sätze, weshalb MÜ zunehmend auch negative Implikationen für den Fremdsprachenerwerb haben.

Neben der sprachinternen Komplexität des Deutschen und Chinesischen und ihrer großen sprachlichen sowie kulturellen Unterschiede stellt auch die Interlingualsprache – meistens Englisch – eine große Barriere dar. Es gibt zurzeit wenige deutsch-chinesische Parallelkorpora. In vielen Fällen werden die korpus-basierten Übersetzungsergebnisse über Englisch erzeugt (vgl. Bastin 2020). So betrifft das Alignment der deutschen und chinesischen Wortbildungen im Prinzip die Bedeutungs- und Grammatiküberschneidung von drei Sprachenpaaren: Deutsch-Chinesisch, Deutsch-Englisch und Englisch-Chinesisch.

Die DE-ZH-MÜ ist daher zurzeit theoretisch, technisch und qualitativ von der EN-ZH-MÜ abhängig. Wegen der Sprachunterschiede der drei Sprachenpaare treten zusätzliche Probleme und Ambiguitäten in DE-ZH-MÜ auf. Für Lexika- und MÜ-Anbieter ist es daher nötig abzuwägen, ob es sich aus ökonomischer und qualitativer Perspektive lohnt, mehr parallele, bilinguale deutsch-chinesische Sprachdaten anzuwenden.

1.2 Forschungsfragen dieser Arbeit

Das schriftliche Alignment der Wortbildungen des Deutschen und Chinesischen ist ein zentrales Forschungsobjekt dieser Arbeit. Zum Prozess des Alignments in den beiden Sprachen sind zuerst drei Fragen zu stellen:

[1] Siehe zu den Grundtheorien des Alignments der Korpora bei Wèi, Lù et al. 2014: 42–67.

1) Wie können die Wortbildungen des Deutschen und Chinesischen segmentiert werden?

In jeder modernen natürlichen Sprache ist es möglich, unendlich viele verschiedene semantische Konzepte auszudrücken. Das gemeingebräuchliche Inventar an Grundmorphemen ist dagegen stets begrenzt. Die Kombination von zwei oder mehreren lexikalischen Einheiten für einen neuen Begriff ist deswegen ein generelles Phänomen. Aber welche Kombinationen als Wörter definiert werden können, ist von Sprache zu Sprache unterschiedlich. Die Definition von Wort ist „uneinheitlich und kontrovers" (Bußmann 2002: 750). Anders formuliert gibt es in konkreten Fällen Unsicherheit, die Worteinheit von den Einheiten des Morphems und der Phrase zu unterscheiden. In modernen alphabetisch-geschriebenen Sprachen gibt es das Spatium, das dem menschlichen und künstlichen Gehirn als Wortgrenze gilt. Ein Kompositum ist in diesem Fall eine Buchstabenfolge, die mithilfe von beiderseitiger Interpunktion einheitlich erscheint und mindestens in zwei weitere lexikalische Morpheme zu zerlegen ist. Im chinesischen und japanischen Schriftsystem ist es problematisch, die Grenze zwischen Morphem/Wort und Wort/Phrase zu erkennen, weil in den Texten Morphem-repräsentierende Schriftzeichen auftreten und keine Spatien zwischen Worteinheiten verwendet werden.

In Kurzform erläutert: Wortbildungen werden in deutschen Korpora zuerst durch Tokenisierung (die Segmentierung eines Texts auf der Wortebene) behoben. Dabei werden die als Wörter gültigen Zeichenfolgen in chinesischen Korpora vor allem durch wörterbuchbasiertes Matching herausgefunden. Konkrete Probleme – wie die Ambiguität der Wortsegmentation in Texten und die lexikalisierten Einheiten in Wörterbüchern – tragen dazu bei, dass die chinesische Wortsegmentation unmöglich das Ziel absoluter Korrektheit erreichen kann.

2) Wie kann das multilinguale Alignment auf verschiedenen sprachlichen Niveaustufen durchgeführt werden?

Die drei sprachlichen Einheiten Grundwort/einfaches Wort, Wortbildung und Phrase können alle für eine eigenständige lexikalische Bedeutung stehen. In welcher Einheit ein neuer Ausdruck am häufigsten auftritt, ist dabei von der jeweiligen Sprache abhängig und unterschiedlich. Im Deutschen dominieren Komposita, wohingegen im Englischen häufiger Phrasen Verwendung finden: „4.2% of the English words which were written separately were written together in German" (Berg 2012: 8). Im Chinesischen erscheint eine Wortbildung (合成词 *héchéngcí* – „komplexes Wort") meist als eine Folge von zwei bis zu über fünf Zeichen. Bspw. ist der physikalische Terminus ‚inertia force' im Englischen eine Phraseneinheit. Das deutsche Äquivalent ‚Trägheitskraft' ist ein Determinativkompositum, wohingegen es sich im Chinesischen um ein dreisilbiges komplexes Wort handelt: 惯性力 *guànxìnglì*.

Damit ein System solche unregistrierten Wortbildungen und Ad-hoc-Bildungen übersetzen kann, müssen Selbstlernfunktionen anhand der vorhandenen Wortbildungen durchgeführt werden. Es ist daher sinnvoll, Wortbildungen des Deutschen nach der Tokenisierung weiter in Morpheme zu zerlegen und Wortbildungsmodelle statistisch zu analysieren (vgl. Popović, Stein und Ney 2006: 8f.). Da im Chinesischen die große Mehrheit (ca. 95%) der Morpheme monosyllabisch ist, also einem einzelnen Schriftzeichen entspricht (vgl. Wang 2019a: 276ff.), kann an einem Wort, das aus mindestens zwei Zeichen besteht, eine Wortbildungsanalyse durchgeführt werden, z. B. zur Kollokationswahrscheinlichkeit der Zeichen in einem Wort oder zu den Wortbildungsalgorithmen.

3) Bei der Bildung der lexikalischen Einheiten divergieren Deutsch und Chinesisch oft auf grammatischer, logischer und kultureller Ebene. Wie kann bilinguales Alignment diese Barriere effektiver überwinden?

Deutsch und Chinesisch haben eine Gemeinsamkeit ob ihres hohen Anteils an Komposita und sonstigen Wortbildungen im Gesamtwortschatz und einer hohen Produktivität für Okkasionalismen. Die Wortbildungsmethoden der beiden Sprachen sind jedoch kompliziert und ihre Unterschiede deutlich größer und weniger erforscht als die zwischen Deutsch und anderen indogermanischen Sprachen. Es gibt deswegen manche zusätzlichen Fehlertypen, die bei der EN-ZH-MÜ nicht auffällig oder nicht vorhanden sind. Welche Arten von Unterschieden die Wortbildungen des Deutschen und Chinesischen haben und welche Fehlertypen deswegen häufiger auftreten, ist daher eine essentielle Forschungsfrage.

In jeder Sprache entwickeln sich jeden Tag neue Wörter und Ad-hoc-Bildungen, die zur Kommunikation gebraucht werden. Der Sprecher einer Sprache kann sie meist anhand seines erworbenen Wortschatzes, der Grammatik und dem Kontext verstehen. Auch Autoübersetzer müssen über diese Kompetenz verfügen. Dies geschieht durch Akquirierung durch Algorithmen anhand bekannter Wortbildungen. So tritt die finale Forschungsfrage hinzu:

4) Welche Algorithmen müssen in Korpora eingesetzt werden, damit die unregistrierten Wortbildungen in höherer Qualität maschinell übersetzt werden?

In den meisten Fällen kann die Semantik eines neugebildeten Kompositums aufgrund der Semantik der Gliedmorpheme richtig geschlussfolgert werden. Um der KI diese Kompetenz beizubringen, bedarf es einerseits statistischer Datenanalysen an Grundmorphemen, Wortbildungsmodellen und Kontexten. Andererseits müssen auch die Wortbildungsgrammatiken und die semantischen Beschreibungen berücksichtigt werden.

Viele multilinguale Autoübersetzer übersetzen, wie oben ausgeführt, zwischen dem Deutschen und Chinesischen durch das Interlingual Englisch. Obwohl Englisch gewisse Vorteile mit sich bringt, als Interlingual verwendet zu werden, vervielfachen sich dadurch die Ambiguitätsfälle. Bspw. wird das

chinesische Wort für *[politische] Rechte* 右派 *yòupài* von Google Translate als *richtig* ausgegeben. Das deutsche Kompositum *Porzellanladen* wird als 中国店 *Zhōngguò diàn* übersetzt (Stand: 21.01.2020), weil das Wort für „Porzellan" im Englischen „*china*" ist. Diese beiden Fehler passieren sowohl im kontextfreien Fall als auch in vielen Satzkontexten. Aus Platzgründen wird im Folgenden zwar nicht auf die so ausgelösten Übersetzungsfehler eingegangen. Werden solche aber zwischen dem Deutschen und Chinesischen (und umgekehrt) thematisiert, so sei darauf hingewiesen, dass dafür häufig das Interlingual Englisch verantwortlich ist.

2 Probleme der Wortbildungssegmentation in einzelsprachlichen Korpora

Die deutschen und chinesischen Wortbildungsprinzipien sind in mancher Hinsicht vergleichbar: Zusammensetzung lexikalischer Morpheme (Komposition) und affixoide Wortbildungen (Derivation) sind in beiden Sprachen grundlegende Bildungstypen. Die chinesische Wortbildung hat zusätzlich einen spezifischen Typ: die Verdopplung von Morphemen/Wörtern. Die größten Unterschiede liegen im Flexionsbewusstsein, den Morphemvarianten und den Wortbildungsgrammatiken.

Die deutsche Wortbildung funktioniert stark über grammatische Beugung, während Chinesisch einen isolierenden Sprachbau ohne jegliche Beugungsformen verwendet. Morpheme im Deutschen unterscheiden sich in Basis- (auch lexikalische), Wortbildungs- (inkl. Präfixe, Zirkumfixe und Suffixe) und Flexionsmorpheme sowie Fugenelemente. Im Gegensatz dazu stellen Basismorpheme im Chinesischen die absolute Mehrheit dar, zu dem sind Wörter nicht flektierbar. Etymologisch stammt ein Großteil der allgemeingebräuchlichen deutschen Morpheme aus dem Lateinischen, Altgriechischen und sonstigen indogermanischen Sprachen, was die Wortbildungsvarianten vervielfacht und die automatische Wortzerlegung erschweren kann. Im Gegensatz dazu machen im Chinesischen autochthone Morpheme den wesentlichen Anteil aus. Die Morpheme werden durch Schriftzeichen repräsentiert und die Wortbildungsstrukturen können im Allgemeinen zeichenbasiert und grammatisch regelmäßiger dargestellt werden. Eine der größten Schwierigkeiten der chinesischen Morpheme ist die Tatsache, dass eine Zeichenform mehreren Morphemen (wie Heteronymen) und Bedeutungen (inkl. Grund-, Erweiterungs- und figurativen Bedeutungen) sowie verschiedenen Wortarten entsprechen kann.

Kontrastiv lässt sich zusammenfassen, dass die deutsche Wortbildung vor allem hinsichtlich ihrer Grammatik und Morphemvielfältigkeit kompliziert ist. Im Chinesischen erschweren einsilbige, dafür aber oft polyseme Morpheme die Wortbildungsanalyse und -segmentierung. Bspw. wurden die

Begriffe 三体 (*sāntǐ*, drei „Körper“) und 三体人 (*sāntǐ rén*, Menschen von *sāntǐ*) in einem chinesischen Sciencefiction-Roman von LIU Cixin erfunden. Dabei kann 体 *tǐ* mehrere Bedeutungen haben, sodass man den Begriff ohne Kontext und Vorkenntnisse nicht verstehen kann. In den deutschen Publikationen wird *sāntǐ* mit ‚die drei Sonnen‘ (so auch der Buchtitel) oder Trisolaris (für alle Ableitungen wie *Trisolaris-Stunden* etc.) übersetzt. ‚Trisolarier‘ – die Bewohner der ‚drei Sonnen‘ – für *sāntǐ rén* setzt sich aus dem lateinischen Präfix *tri-*, dem Lehnmorphem *solar* und dem Flexionsmorphem *-er* zusammen.[2]

Die automatisch erzeugte Übersetzung einer unregistrierten Wortbildung muss also auf der Basis der vorhandenen Wörter sowie Wortbildungsgrammatiken basieren. Segmentationen der Wortbildungen in Korpora sind dafür Voraussetzung.

2.1 Probleme der inneren Zerlegung der deutschen Wortbildungen

Anhand der Schriftform natürlichsprachlicher Wörter kann die Wortsegmentation in drei Fälle untergegliedert werden: 1) alphabetgeschriebene einfache Wörter, die aus einem lexikalischen Morphem (ohne grammatischer Flexion) bestehen; 2) alphabetgeschriebene Wortbildungen, die aus mehreren Morphemen bestehen; 3) chinesische (oder japanische) Wörter ohne Spatium-Markierung (siehe dazu Abschnitt 2.2). Die Wortsegmentation der deutschen Korpora bezieht sich auf den ersten und zweiten Fall. Wörter, die aus zwei oder mehreren Morphemen bestehen, werden in kleinere Bestandteile zerlegt, sodass das maschinelle Lernen solcher Wortbildungen besser durchgeführt werden kann.

Wie erwähnt sind Komposition und Derivation die zwei dominierenden Wortbildungstypen des Deutschen. Derivationsaffixe sind meistens begrenzt und auf bestimmte Wortarten festgelegt, sodass die derivierten Wortbildungen (mit Ausnahme der entlehnten bedeutungstragenden Affixe) im Natural Language Processing (NLP) in den meisten Fällen morphosyntaktisch erkannt und analysiert werden können. Außerdem können die Techniken für die Segmentation der Derivationswörter für EN-ZH-MÜ grundsätzlich für DE-ZH-MÜ übernommen werden. Im Vergleich dazu sind die Kompositionsfälle im Deutschen nicht nur häufiger als im Englischen, sondern auch aufgrund ihrer Wortbildungsstruktur oft komplizierter zu analysieren und stärker von statistischen Analysen abhängig, was mehr Probleme bei der maschinellen Wortanalyse und der MÜ verursachen kann. Aus diesem Grund wird in diesem Kapitel die Segmentation der Komposita in Morpheme stärker

[2] Vgl. dazu LIU Cixins „Trisolaris-Trilogie“ (刘慈欣：《三体三部曲》 2007–2010), übersetzt ins Deutsche von Martina Hasse und Karin Betz 2017–2019.

berücksichtigt. Im komplizierten Fall besitzt ein deutsches Wort alle vier Morphemtypen, z. B. *Schönheitsoperationen*: Grund- und freies Morphem: *schön*; gebundenes Morphem: *operātio* [lat.]; Derivationsmorpheme: *-heit* und *-tion*; Fugenelement: *-s-*; Flexionsmorphem: *-en*. Die Zerlegung von Komposita ist in diesem Kapitel daher Schwerpunkt der Wortsegmentation.

Ein Kompositum besteht aus zwei Konstituenten, wobei eine Konstituente in manchen Fällen in untergeordnete Konstituenten weiter zerlegt werden kann. Dazwischen können Fugenelemente vorkommen, wie z. B. bei *Tischlampenschirme*. Die erste Konstituente ist weiter zerlegbar in *Tisch* und *Lampe*. Die Kombination von der ersten Konstituente und dem Fugenlaut entspricht in vielen Fällen einer Flexionsform, aber nicht immer. Da die deutschsprachige Wortsegmentation fast immer vorwärts (also vom ersten Buchstaben an) durchgeführt wird, ist eine korrekte Kompositumzerlegung nur möglich, wenn das erste Wortglied erkannt wird (vgl. Langer 1998: 1). Für eine automatische Aufteilung der Komposita sind morphosyntaktische Analysen und eine Fugenelementeliste nötig.

Tischlampenschirm legt exemplarisch ein weiteres Problem der Kompositumzerlegung offen, nämlich den Fall, dass ein mittiges Glied sowohl mit dem vorderen als auch dem hinteren Glied ein sinnvolles Wort bilden kann: in diesem Fall also *Tischlampe* und *Lampenschirm*. Ob die KI in solchen Fällen eine hierarchisch korrekte Aufteilung vornehmen kann und zu einer semantischen Analyse gelangt, entscheidet direkt über die Übersetzungskorrektheit.

Zusammenfassend muss die Kompositumzerlegung mindestens drei Punkte erfüllen: 1) die hierarchisch konstituierte Segmentation; 2) die Bestimmung der Grund- und peripheren Morpheme verschiedener Typen im Wort, inkl. Wortwurzel, Wortbildungs-, Flexionsmorpheme und Fugenelemente; 3) das *Part-of-Speech-Tagging* (POS-*Tagging*) der Grundmorpheme (vgl. auch Trommer 2010: 262 sowie grundlegend Zhōu und Duàn 1999). Mit den Zerlegungsergebnissen der Komposita ist es weiterhin möglich, algorithmische Analysen bei Wortbildungsstrukturen anzugehen, z. B.: a) wie wahrscheinlich kommt ein Fugenelement nach einer bestimmten Buchstabenfolge vor; b) wie wahrscheinlich hängen die beiden Konstituenten im Wort und im Satzkontext zusammen; c) im Fall, dass Parallelkorpora begründet werden, kann auch berechnet werden, welches Flexionsmorphem des Deutschen häufiger mit einem Wort/Morphem des Chinesischen korreliert.

Der Prozess der Kompositazerlegung wird im Englischen „decomposition“ genannt. Die konkreten Anwendungsverfahren zeigen bspw. Sugisaki und Tugger (2018: 142f.). Ihrer Studie nach passieren bei drei Kompositionsarten häufig Segmentationsfehler: bei fachlichen Termini (wie *Doldensurre*), bei Fremdwörtern (wie *Backstage*) und bei aus nativen und fremdsprachlichen Morphemen zusammengesetzten Komposita (wie *Bioei*).

2.2 Schwierigkeiten bei der Segmentation chinesischer komplexer Wörter

Die Segmentation des Chinesischen verläuft in entgegengesetzter Richtung zu der des Deutschen: Segmentationseinheiten werden in elektronischen Wörterbüchern eingetragen und durch Matching als Wörter segmentiert. Eine Segmentationseinheit kann minimal aus einem, jedoch auch aus über fünf Schriftzeichen bestehen und entspricht nicht unbedingt einem Wort. Für die Segmentation einer Zeichenfolge in Segmentationseinheiten wurden viele verschiedene Methoden entwickelt, wobei zur Verarbeitung eines chinesischen Korpus im Sinne einer höheren Präzision oft mehrere Methoden hybridisch eingesetzt werden (vgl. Wang 2019a: 251f.).

Da die meisten Schriftzeichen einsilbige Morpheme repräsentieren und die bi- sowie polysyllabischen Morpheme in den meisten Fällen zu den Lexemen eines Wörterbuchs gehören, kann die KI die lexikalischen Grundeinheiten meist problemlos erkennen. Herausfordernder ist, innerhalb einer Zeichenfolge ohne Interpunktion die Wortgrenzen richtig zu erkennen. Im Deutschen ist die innere Wortzerlegung sinnvoll, um das maschinelle Lernen von Wortbildungen zu verbessern. Im Gegenteil dazu ist im Chinesischen die Wortsegmentation für fast alle Arten der Textverarbeitung obligatorisch, weil ohne eine korrekte Segmentation weitere Analysen von Phrasen und Sätzen sowie richtige Übersetzungsergebnisse unmöglich sind (Hóu 1999: 98ff.).

Wegen der hohen Wahrscheinlichkeit, dass ein Zeichen einer komplexeren Wortbildung angehört, in Ermangelung von Spatien und nicht zuletzt aufgrund undeutlicher Grenzen zwischen den Einheiten Morphem/ Wort sowie Wort/ Phrase ist die Wortsegmentation im Chinesischen ambig und stellt eine Herausforderung dar. Bspw. kann die Zeichenfolge 中国人民生活 *Zhōngguó rénmín shēnghuó* nach dem Matching mit der Wörterliste in mehrere Alternativen segmentiert werden. Jedes der sechs Zeichen kann ein eigenständiges Wort darstellen. Durch jeweils zwei benachbarte Zeichen lassen sich fünf bisyllabische Wörter bilden: 中国 *Zhōngguó*, 人民 *rénmín*, 生活 *shēnghuó*, 国人 *guórén* und 民生 *mínshēng*. Zusätzlich gibt es auch eine trisyllabische Kollokation: 中国人 *Zhōngguórén*. Ohne eine Analyse hat diese Zeichenfolge also 16 Segmentierungsmöglichkeiten. Das Matching einer Zeichenfolge mit einer Wörterdatenbank reicht deswegen bei weitem nicht aus, um in komplizierten Fällen eine korrekte Segmentation durchzuführen.

Es gibt mehrere Varianten für maschinelle Segmentationsmethoden in chinesischen Korpora. Darunter sind die sog. ‚maximalen Matchingmethoden' (最大匹配法 *zuì dà pǐpèifǎ*; inkl. Vorwärts- und Rückwärtsmatching) die ältesten und grundlegendsten. Funktionsprinzipien und Arbeitsprozesse verschiedener Segmentationsmethoden und die Methoden zur Disambiguierung der Segmentation werden in Wang (2019a: 254–258), Hóu (1999: 100–

117) und Wáng (2005: 35–39) analysiert. Das obige Beispiel kann mit der statistik-basierten Wortvernetzungsmethode – zusammengefasst – wie folgt durchgeführt werden: Nachdem alle möglichen Wörter erkannt und in Wortvernetzungsform begründet wurden, werden einstellige Algorithmen (bezogen auf ein einzelnes mögliches Wort) sowie zweistellige Algorithmen (bei zwei benachbarten möglichen Wörtern) durchgeführt. Die Segmentationsvariante mit dem besten ‚arg max'-Ergebnis (Argument des Maximums) wird als die optimale Variante ausgewählt. Im Fall der Beispielphrase 中国/人民/生活 *Zhōngguó/rénmín/shēnghuó* (*China / Volk / Leben*) ist das die semantische Bedeutung ‚*das Leben des chinesischen Volks*'.

Tests zufolge hat die Segmentation bis zum Ende des 20. Jahrhunderts bereits eine durchschnittliche Präzision von 89,4% erzielt (vgl. Hóu 1999: 105). Präzision und Effektivität zugleich auf einem akzeptablen Niveau zu verbessern, ist dabei das Hauptziel. Bei der statistik-basierten Wortvernetzungsmethode (vgl. obiger Absatz) muss die KI vergleichsweise komplexe Algorithmen ausführen: Als Vorphase einer MÜ verlangsamt das wiederum den Gesamtprozess (vgl. Liú et al. 2009: 124).

Welche Zeichenfolgen segmentiert werden können, ist von dem jeweiligen zugrundeliegenden elektronischen Wörterbuch abhängig, wobei die eingetragenen Lexeme Segmentationseinheiten sind. Welche Wörter/Phrasen als Segmentationseinheiten definiert werden, ist auch von dem Verarbeitungsziel abhängig. Im Fall der semantischen Analysen ist es z. B. sinnvoll, eine Zeichenfolge, deren Bedeutung nicht direkt durch ihre Bestandsmorpheme/-wörter zu analysieren ist und nicht trennbar gebraucht wird, als eine lexikalische Einheit zu definieren. Im umgekehrten Fall ist das optional. Bspw. macht es keinen großen Unterschied, die Zeichenfolge ‚中国人' *Zhōngguórén* als ein Wort oder eine Phrase zu definieren und zu verarbeiten. Denn einerseits ist es möglich, mithilfe der Wort-/Phrasenbildungsgrammatik die semantische Bedeutung für ‚Menschen aus China' zu bilden. Andererseits kann zwischen einer Zeichenfolge die Partikel 的 *de* oder ein Attributmorphem (wie 老 *lǎo* „alt", 女 *nǚ* „weiblich") vorkommen, was keine semantische Änderung der Zeichenfolge mit sich bringt (vgl. Féng 2001: 26f.).

Als Gegensatz zu diesem Beispiel ist es nötig, den recht neuen Internetausdruck 吃瓜群众 *chīguā qúnzhòng* als eine Segmentationseinheit einzutragen, obwohl er in Korpora relativ selten vorkommt (‚*Schaulustige*' bzw. ‚*das Melonen essende Volk*' [wörtliche Übersetzung]. Mit diesem Ausdruck werden Leute bezeichnet, die keine Ahnung von einer Sache haben, aber dabei zuschauen und Irrelevantes äußern.) In diesem Fall ist es der KI unmöglich, durch die Phrasenstrukturgrammatik die richtige Semantik anhand der Bestandswörter auszugeben. Wenn 的 *de* innerhalb der beiden Bestandswörter eingefügt wird, bleibt die Bedeutung prinzipiell unverändert.

Wurde ein komplexes Wort erkannt, sind auch Analysen der Konstituentenstruktur nötig. Da ein einzelnes Zeichen mit hoher Wahrscheinlichkeit ein einsilbiges Morphem repräsentiert, sind diese Analysen meist zeichenbasiert (vgl. Xú 2008: 103).

Kollokationswahrscheinlichkeitsalgorithmen beziehen nicht nur die Bestandszeichen/-wörter eines komplexen Wortes, sondern auch die miteinander korrelierenden Wörter/Zeichen einer Phrase mit ein. Bspw. ist es in den Phrasen 中国的人 *Zhōngguó de rén* und 中国老人 *Zhōngguó lǎorén* ebenfalls notwendig, die Korrespondenz von 人 zu 中国 (*rén* zu *Zhōngguó*) trotz des zusätzlichen Zwischenglieds zu erkennen und es als einen Teil von P (人 *rén* |中国 *Zhōngguó*) zu zählen. P (B|A) repräsentiert die bedingte Wahrscheinlichkeit von B durch A, die mit der Bayes-Theorem-Formel errechnet werden kann (in diesem Fall A: 中国 *Zhōngguó* und B: 人 *rén*; vgl. Wang 2019a: 277).

3 Bilinguales Alignment und die MÜ unbekannter Wortbildungen

In der Einleitung wurde skizziert, dass das Alignment des Deutschen und des Chinesischen wegen verschiedener sprachlicher Einheiten, semantischer Ambiguitäten und der schriftlichen, grammatischen und logischen Unterschiede problematisch durchzuführen ist. Die Probleme dahinter sind variantenreich und vielseitig. Aus Platzgründen konzentriert sich dieser Artikel darauf, wie das bilinguale Alignment auf Wort- und Morphemebene auszuführen ist und wie passende Wortübersetzungen erzeugt werden können.

Die vorgestellten Techniken in diesem Kapitel können in verschiedenen Kontexten der korpusbasierten Textverarbeitung Anwendung finden. Teile davon werden bei der EN-ZH-MÜ schon angewendet. Auf Basis der vorhandenen Techniken habe ich eine Analyse durchgeführt und ein Konzept entworfen, welche Methoden für eine bessere und qualitativ höherwertige DE-ZH-MÜ von Wortbildungen angewendet werden können.

3.1 Das Alignment und die MÜ terminologischer Bildungen

Terminologische Wortbildungen werden aus den folgenden Gründen zuerst diskutiert: Erstens haben Termini relativ hohe konzeptionelle Übereinstimmungen in verschiedenen Sprachen, sodass zwischen ihnen interlingual relativ häufig eindeutige Alignments gebildet werden können. Zweitens dient die Setzung von Termini der gezielten Wissensverbreitung und ist daher artifizieller als natürlichsprachliche Äquivalente, die variabler und mit mehr Ambiguitäten aufwarten. Die sprachlich und schriftlich bedingten Wortbildungsunterschiede können deswegen deutlicher analysiert werden. Drittens sind ein

Großteil aller neuen Wörter, die tagtäglich erfunden werden, Termini, weshalb ihre MÜ aus qualitativen und quantitativen Gesichtspunkten eine hohe Effizienz beansprucht.

3.1.1 Die Übersetzung terminologischer Bildungen im Allgemeinen

Anhand der Übereinstimmungsweise der Morpheme können vier Termini-Gruppen als Beispiele genannt werden:

1) Termini, in denen jedes einzelne Basismorphem Entsprechung findet und in identischer Reihenfolge geordnet wird, wie z. B. „Energieerhaltungssatz“ – 能量守恒定律 *néngliàng shǒuhéng dìnglǜ* (Energie – 能量 *néngliàng*, Erhaltung – 守恒 *shǒuhéng*, Satz – 定律 *dìnglǜ*);

2) Termini, in denen die Morpheme in anderer Reihenfolge geordnet werden, wie z. B. Kohlendioxid – 二氧化碳 *èr yǎnghuà tàn* (Kohlen – 碳 *tàn*, di- [lat.] – 二 *èr*, ox- [gri.] – 氧 *yǎng*, -id [gri.] –化 *huà* [Morphem für chemische Umwandlung]);

3) Termini, die sich im Deutschen/Englischen und Chinesischen phonetisch entsprechen, aber im Chinesischen nicht in bedeutungstragende Morpheme zerlegbar sind, wie z. B. Remdesivir – 瑞德西韦 *ruìdéxīwéi* (Medikament gegen das Coronavirus);

4) Termini, in denen die Basismorpheme nur teilweise übereinstimmen, wie z. B. Photosynthese – 光合作用 *guānghé zuòyòng* (photo [lat.] – 光 *guāng*, synthese [gri.] –合 hé, [Funktion] – 作用 *zuòyòng*), oder Fünfeck – 五边形 *wǔbiānxíng* (fünf – 五 *wǔ*, Ec - 边 *biān* [Seite(!)], [Form] – 形 *xíng*).

Bei manchen terminologischen Klassen gibt es die Möglichkeit, anhand spezieller Sprachmodelle die korrekte Übersetzung zu erzeugen. Bspw. gelten bei den geometrischen Zeichnungen die alignierten Sprachmodelle „x-*Eck*“ – „x 边形 *biānxíng*“ (x wird im Deutschen und Chinesischen mit entsprechendem Grundzahlwort gelesen; ungültig bei Dreiecken [三角形 *sānjiǎoxíng*]). Solche Sprachmodelle mit Eins-zu-eins-Entsprechung gelten auch bei den chemischen Verbindungen. Sie können zwar eine präzise und schnelle Übersetzung anbieten, können aber nur bei einem kleinen Teil der Termini angewendet werden. Bei den sonstigen können String-Matching-Algorithmen (auch Ähnlichkeitsalgorithmen genannt) ausgeführt werden, um auf der Basis eines ähnlichen und bekannten Terminus einen unbekannten zu übersetzen (vgl. Wáng et al. 2015: 67f.). Das ähnliche Wort muss mindestens ein identisches Morphem an derselben Wortposition wie das gematchte Wort haben, und die nicht übereinstimmenden Morpheme müssen teils/komplett von derselben Wortart sein.

Für die Festlegung eines Terminus in jeder Sprache gilt es zwischen der deutlicheren Verständlichkeit (Bedeutungsübertragung) und der Beibehaltung eines Konzeptes der Ausgangssprache (phonetische Übertragung oder direkte Übernahme der Wortform) abzuwägen. Auch Semantik-Phonetik-Hybride sind relativ häufig. Allgemein betrachtet wird im Deutschen die wörtliche/phonetische Übernahme bevorzugt, während im Chinesischen die Wahrscheinlichkeit einer terminologischen ‚Lokalisierung' höher ist. Das ist sowohl naturwissenschaftlich/kulturell als auch sprachlich/schriftlich bedingt. In Europa gibt es seit den Anfängen der Wissenschaft die Konvention, Morpheme aus dem Lateinischen/Griechischen etc. für terminologische Bildungen zu übernehmen. Die Entwicklung der modernen Naturwissenschaften wurde vom kulturellen Austausch zwischen den europäischen Ländern begleitet, für die sich eine wörtliche/phonetische Übernahme als effizienter erwiesen hat. Wegen der sprachlichen Gemeinsamkeiten und gleichartiger Schriftsysteme ist eine solche terminologische Übernahme relativ einfach durchzuführen.

Der Austausch zwischen chinesischen und westlichen Sprachen intensivierte sich erst im Verlauf des 19. Jahrhunderts, wobei die polysyllabischen Termini einer europäischen Sprache nur mit Problemen in die chinesische Schrift zu übertragen sind. Wegen der Kategorie Tonsprache und der ideografischen Schrift gibt es im Chinesischen anteilig mehr monosyllabische Morpheme/Wörter von relativ einfacher Silbenstruktur. Es ist außerdem sprachlich konventionalisiert, ein Morphem phonetisch monosyllabisch und schriftlich mit einem Zeichen zu repräsentieren (vgl. Wang 2019b: 44f.). So bevorzugen Chinesen oder Japaner, für wissenschaftliche Konzepte neue Wörter anhand vorhandener Morpheme/Zeichen zu bilden.

Es ist zwar unwahrscheinlich, dass eine KI so kreativ wie menschliche Übersetzer*innen neue Termini erfinden kann, sie kann aber anhand ihrer Selbstlernfunktionen aus multilingualen, nach Taxonomien geordneten Terminologiedatenbanken oder kurzen terminologischen Beschreibungen möglichst verständliche und semantisch passende Übersetzungen anbieten.

3.1.2 Die morphembasierte MÜ terminologischer Bildungen

In Kapitel 2 wurden Wortsegmentierungsmethoden vorgestellt. Um der KI die Übersetzung eines unbekannten Terminus beizubringen, müssen deutsche und chinesische terminologische Bildungen der Korpora zuerst in Morpheme segmentiert werden. Semantisch-phonetisch-hybrid gebildete Termini werden in ihre phonetischen und ihre bedeutungstragenden Bestandteile zerlegt. Die Verarbeitung von Termini mit phonetisch-entlehnten Morphemen wird im folgenden Absatz diskutiert. Dabei kommt die Methode des sog. Fuzzy-Matching zum Tragen, bei dem die alignierten Termini der beiden involvier-

ten Sprachen hinsichtlich ihrer einzelnen Morpheme mithilfe des bilingualen Wörterbuchs abgeglichen werden.

Weiterhin ist es erforderlich zu errechnen, wie wahrscheinlich ein bilinguales Morphempaar in bestimmten Fachgebieten miteinander korreliert. Die Algorithmen gegenseitiger Korrelation heißen mutuale Korrespondenz (aus dem Englischen: *mutual correspondence*; zur Grundidee und Formel vgl. Altenberg 1999: 254). Wenn man einen unbekannten Terminus automatisch übersetzen lassen möchte, muss das System (anhand des Kontexts, so vorhanden) das dazugehörige Fachgebiet erkennen und den Terminus in seine Morpheme segmentieren. Innerhalb des Fachgebiets kann mithilfe der String-Matching-Algorithmen der Terminus mit der größten formellen Ähnlichkeit als Vorbild für das unbekannte Wort recherchiert werden. Anhand der vorhandenen Übersetzung des „Vorbild-Terminus", der mutualen Korrespondenz und der passenden Wortbildungsgrammatik, kann eine logische Übersetzung angeboten werden.

Ein Beispiel liefert der unbekannte Terminus 宇宙质量守恒定律 *yǔzhòu zhìliàng shǒuhéng dìnglǜ*: Anhand (eines Kontextes, so vorhanden, und) der vier letzten Zeichen wird der Fachbereich *Physik* präsupponiert. Nach der Recherche ist der Terminus 能量守恒定律 *néngliàng shǒuhéng dìnglǜ* mit seinen vier identischen Zeichen am Wortende am ähnlichsten (deutsch: *Energieerhaltungssatz*). Wegen der großen Ähnlichkeit werden das deutsche Alignment und die Wortbildungsgrammatik des Vorbild-Terminus übernommen. Letzterer wird im Deutschen als Determinativkompositum gebildet, dessen Bestandsmorpheme in identischer Reihenfolge wie im Chinesischen angeordnet sind. So wird ‚*-erhaltungssatz*' als Kompositionskopf eingesetzt, während die Übersetzungen der beiden determinativen Teile in identischer Reihenfolge vorne beigefügt werden. In der Physik wird 宇宙 *yǔzhòu* am wahrscheinlichsten mit *Kosmos* aligniert; 质量 *zhìliàng* stimmt am häufigsten mit *Masse* überein. Nach statistischen Analysen der zerlegten Komposita findet das System heraus, dass *Masse* als erstem Konstituenten immer ein *-n-* als Fugenelement folgt. Der Übersetzungsvorschlag lautet sodann: *Kosmosmassenerhaltungssatz*.

3.1.3 Die phonetikbasierte MÜ terminologischer Bildungen

Die phonetikbasierte MÜ wird in anderer Weise ausgeführt als die morphembasierte. Zu Vereinfachungszwecken werden nur die phonetischen Übertragungen aus Englisch/Deutsch ins Chinesische diskutiert. Bei chinesischen Termini in Korpora wird ein polysyllabischer Bestandteil als phonetisch-entlehnt erkannt, wenn unter ihnen kein bi- oder polysyllabisches Wort gefunden wird und die Kollokationswahrscheinlichkeit der benachbarten Zeichen sehr gering ist. Daraufhin wird eine Segmentation in Silben und/oder

Phoneme im Deutschen/Englischen und das Matching mit dem chinesischen Zeichen sowie den Pinyin-Silben durchgeführt. Wegen der großen Unterschiede bei den Silbenstrukturen gibt es oft Konsonanten des Deutschen/Englischen, die im Chinesischen nicht übertragen oder zu selbstständigen Silben erweitert werden. Algorithmen über die mutuale Korrespondenz zwischen deutschen/englischen und chinesischen Silben können dann berechnet werden. Wenn der unbekannte Terminus aus phonetischen Morphemen besteht, die dem System bereits bekannt sind, setzt das System die phonetischen Alignments zusammen.

Die Verarbeitungsweise der phonetischen Morpheme unterscheidet sich bei Eigennamen und Nicht-Eigennamen. Eine Liste der häufigsten Zeichen eingetragener Personen- und Ortsnamen außerhalb des CJKV-Kulturkreises[3] kann statistik-basiert erstellt und für die Übersetzung unregistrierter Eigennamen verwendet werden (vgl. Hóu 1999: 123). Hier ist außerdem zu beachten, dass die phonetischen Übertragungen westlicher Eigennamen ins Chinesische (wegen der kulturellen, dialektalen/sprachlichen Unterschiede und verschiedenen Standards der Transkription) in Festlandchina, Taiwan und Hongkong/Macau relativ weit differieren können. Von Vorteil wäre es daher, Zeichenlisten der Eigennamen regional differenziert zu erstellen, abzugleichen und anzuwenden. Wenn das phonetische Morphem nicht der Eigennamenliste angehört und die Kontextanalyse einen Eigennamen ausschließt, kann es anhand der Taxonomie und kurzer Begriffsbeschreibungen weiter analysiert werden.

Wegen der ideografischen Eigenschaft der chinesischen Schriftzeichen beeinflusst auch die Zeichenform von Fremdwörtern mehr oder weniger das chinesische Verständnis bei von Hand angefertigten Übersetzungen: Unter den ähnlich klingenden Zeichen wird jenes bevorzugt, dessen Bedeutung dem zu beschreibenden Konzept am ehesten entspricht. Damit aber eine KI passende Schriftzeichen automatisch auswählen kann, wäre eine nach semantischen Radikalen und Pinyin angeordnete Zeichendatenbank vonnöten. Bestimmte Fachgebiete/Kategorien/Objektstatus könnten mit einer oder mehreren bevorzugten Radikalen verknüpft werden, bspw. 木 *mù* und 艹 *ǎo* bzw. *cǎo* für Botanik, 氵 *shuǐ* für flüssige Objekte, 口 *kǒu* für mit dem Mund eingenommene Medikamente etc. Wenn ein unregistrierter Terminus phonetisch ins Chinesische übertragen werden soll, ist anzunehmen, dass sich die KI auf Grundlage der Fakten der phonetischen mutualen Korrespondenz sowie der Zeichengebrauchshäufigkeit für Fremdwörter und der bevorzugten Radikale der Begriffskategorie für die optimalen Zeichen entscheidet.

[3] CJKV steht für China, Japan, Korea und Vietnam. Die Abkürzung bezieht sich auf den vom Konfuzianismus, von der chinesischen Schrift etc. beeinflussten ostasiatischen Kulturkreis.

3.2 Disambiguierung mithilfe grammatischer Regeln

Die Verwendung des ähnlichsten eingetragenen Wortes, das mit Ähnlichkeitsalgorithmen recherchiert wird, ist theoretisch die effizienteste Methode zur Übersetzung unbekannter Wortbildungen. Die im letzten Kapitel vorgeführten Methoden für terminologische Bildungen können auch für viele allgemeine Wortbildungsfälle angewendet werden. Eine regelbasierte Überprüfungsfunktion ist in solchen Fällen aber umso nötiger, um Ambiguitäten der Ausgangs-, Übergangs- und Zielsprache bei der MÜ zu reduzieren.

3.2.1 Die regelbasierte MÜ-Überprüfung

An dieser Stelle führe ich ein chinesisches Kompositum nebst falscher Autoübersetzung und falscher Pinyin-Umschrift an: 藏猫猫 *cángmāo māo* – **tibetische Katze* / **tibetische Katze und Katze*.[4]

Die größte Schwierigkeit dieser Übersetzungseinheit liegt beim Heteronym 藏: *cáng – verstecken* und *zàng – Schatz*/*Tibet*. Da im MÜ-System phonetische Umschrift und Übersetzung miteinander unverbunden durchgeführt werden, bietet es die korrekte Zeichenaussprache *cáng* (wegen der höheren Wahrscheinlichkeit dieser Lesart), aber die falsche Zeichenübersetzung *Tibet* im Wortkontext an. (Grund dafür ist die hohe Frequenz der Wortform „藏 *zàng* + *Tier*" für tibetische Tierarten.) Da 藏猫猫 *cáng/zàng māomao* nicht in die bilingualen Wörterdatenbanken eingetragen wurde und nicht durch ein ähnlich strukturiertes vorhandenes Wort analysierbar ist, sind grammatikbasierte Analysen für eine korrekte MÜ obligatorisch. Das System muss anhand der beiden heteronymen Alternativen zwei Wortsegmentationsmöglichkeiten sowie POS-Tagging miteinberechnen: 藏 *cáng*/v 猫猫 *māomao*/n und 藏猫 *zàngmāo*/n 猫 *māo*/n. Die Alternative im ersten Fall ist nach der Verb-Objekt-Grammatik gültig. Im Gegensatz dazu ist zwar die erste Konstituente der zweiten Alternative gültig, aber die ABB-Struktur ist ungültig, wenn A und B beide nominal sind (vgl. Wang 2019a: 275). Mit den genannten Analysen kann die zweite Möglichkeit ausgeschlossen und die richtige Pinyin-Umschrift *cáng māomao* sowie eine besser passende deutsche Übersetzung (wie *die Katze verstecken*) angeboten werden. Aber die korrekte Übersetzung – *das Versteckspiel* – kann theoretisch nur anhand eines Satzkontexts ermittelt werden.

[4] Getestet durch Eingabe als einzelnes kontextfreies Wort in Google Translate und Baidu Fanyi (Stand: 20.02.2020; in der englischen Übergangssprache sind dieselben Fehler aufgetreten).

3.2.2 Sonstige Probleme der morphosyntaktischen und semantischen Analysen

Wegen der großen sprachlichen Entfernung gibt es immer noch viele unbekannte Wortbildungsfälle, die nicht durch String-Matching-Algorithmen und einfache grammatikbasierte Überprüfungsfunktionen zu übersetzen sind. Bspw. ist 肉食者 *ròushízhě* nicht in den Datenbanken eingetragen. Das ähnlichste Wort 素食者 *sùshízhě* entspricht dem deutschen *Vegetarier*, dessen Struktur für das unbekannte Wort nicht anwendbar ist. Das liegt daran, dass beide Wörter im Chinesischen aus einem prädikativen Grundwort und dem Suffix 者 *zhě – jemand –* bestehen, wohingegen im Deutschen das Wortbildungselement ‚-er' obligatorisch ist.[5] Um von 肉食者 *ròushízhě* auf das deutsche Äquivalent *Fleischesser(-in)* zu kommen und umgekehrt, hat ein MÜ-System diverse Schwierigkeiten zu überwinden.

Erstens ist das Zeichen 食 *shí* eine multi-kategoriale Wortart und kann als Verb wie auch als Nomen fungieren. Das Wort 肉食 ist im modernen Chinesisch häufiger substantivisch (*Fleischprodukt*) als verbal (*Fleisch essen/fressen*), denn im verbalen Fall steht das Objekt vorne, was der gebräuchlicheren Reihenfolge (SVO) im modernen Standardchinesisch widerspricht.

Zweitens sind die Bestandsmorpheme der Komposita hierarchisch anders geteilt: im Chinesischen 肉 食 | 者 *ròushí|zhě* und im Deutschen Fleisch|esser.

Drittens gibt es zwischen 食 *shí* und *essen/fressen* eine Bedeutungsüberschneidung. Jeder, der einen Mund hat, kann 食 *shí* als Handlung ausführen. Erst das Suffix 者 *zhě* schränkt ein, dass es sich um einen Menschen handeln muss. Wenn statt 者 *zhě* das Wort für *Tier* 动物 *dòngwù* auftaucht, entspricht es dem deutschen Wort *Fleischfresser*. Im Deutschen markieren das die unterschiedlichen Wortwurzeln ‚-ess-' und ‚-fress-'. Sowohl Menschen als auch andere Lebewesen und Nicht-Lebewesen – soweit sie maskuline Substantive darstellen – können eine ‚-er'-Endung nach einem verbalen Wortstamm tra-

[5] Im Chinesischen fand das Wort 肉食者 *ròushízhě* bereits im „Zuozhuan" (ca. 375–351 v. Chr. von ZUO Qiuming (左丘明] verfasst) Verwendung und meint die (Fleisch essenden) Leute der herrschenden Sozialschicht ("肉食者鄙，未能远谋" *„ròushízhě bǐ, wèi néng yuǎn mó*, *Die Leute an der Macht sind kurzsichtig und können keinen langfristigen Plan machen"*, 《左传·曹刿论战》 *„Zuǒzhuàn, Cáo Guì lùn zhàn"*). Das Wort 素食 *sùshí* existierte auch schon zu jener Zeit und beschrieb ungekochtes pflanzliches Essen. Da im Altchinesischen die Objekt-Verb-Folge ein häufiges grammatisches Phänomen war, steht in 肉食 *ròushí* / 素食 *sùshí* das Objekt vorne, was für das moderne Chinesisch untypisch ist. Als der Vegetarismus sich im 20. Jhdt. in China verbreitete, wurde anhand von 素食 *sùshí* das neue Wort 素食者 *sùshízhě* gebildet, mit dem Antonym 肉食者 *ròushízhě*. Im Deutschen wurde *Vegetarier* vom englischen *vegetarian* entlehnt. Das determinativ-kompositorisch gebildete Wort *Fleischesser* trat dann später auf.

gen, wie z. B. Wasserkocher (烧水壶 *shāoshuǐhú*), Pflanzenfresser (食草动物 *shícǎodòngwù*) etc.

Viertens muss ein deutsches Nomen mit den Modellen der analytischen sowie der generativen Morphologie über Genus, Kasus und Numerus verarbeitet werden (zur Methodik der generativen Morphologie vgl. Trommer 2010: 240–262). Die Flexionen des Deutschen werden im Chinesischen durch manche Grundwörter/-morpheme oder Wortfolgen ausgedrückt, bspw. ‚-in' – 女 *nǚ* (menschlich) / 母 *mǔ* (manche Tiere) / 雌 *cí* (biologisches Femininum); / Hühnerfutter / die Hühner etw. fressen – 鸡食 *jīshí*, Hühner essen – 食鸡 *shíjī*.

Damit die KI in diesem Fall eine korrekte MÜ anbieten kann, müsste sie die nachstehenden sechs Punkte erfüllen. Zuvor sei kurz auf den gegenwärtigen Stand der Technik verwiesen: Führt man entsprechende Tests bei MÜ-Systemen wie Google oder Baidu durch, kann nur gemutmaßt werden, ob das interne System die dargestellten Prozesse bereits durchlaufen lässt bzw. auf wie viele sprachliche Informationen zugegriffen wird. Beschrieben wird nachstehend eine Art Optimalzustand aus Sicht der zu erwartenden Übersetzungsqualität.

(1) Manche untypischen, aber gültigen grammatischen Regeln gilt es einzutragen. Im Beispiel 肉食者 *ròushízhě* müsste dann auch die OV-Struktur erkannt werden.

(2) Hierarchische Zerlegungen in jedes kleinste Morphem und dessen POS-*Tagging* sind vonnöten, wie die Beispiele „[肉 *ròu*/$_{n}$|食 *shí*/$_{v/n}$]/$_{v/n}$||者 *zhě*/$_{n}$" und „Fleisch/$_{n}$ || [ess/$_{bm}$ | er/$_{fm}$]/$_{n}$" zeigen.[6]

(3) Es muss mithilfe von sprachlichen Regeln disambiguiert werden; im genannten Beispiel sind die Zeichengebrauchsregeln von 者 *zhě* entscheidend. Das Suffix 者 *zhě* kann mit einem Verb, Adjektiv, Nomen oder einer Numerale ein komplexes Wort bilden, aber unter den Nomina sind nur die mit 工作 *gōngzuò* oder 主义 *zhǔyì* endenden möglich. Da das Nomen 肉食 *ròushí* nicht dem nominalen Fall entspricht, muss es als Verb erkannt werden.

(4) Es muss ein Fuzzy-Matching zwischen den interlingualen Sprachmodellen „Verbaldeterminativ + Nominalkopf/者 *zhě*" des Chinesischen und „Verbalstamm + *-er*" des Deutschen begründet werden. Das gilt auch, wenn Verbalmorphem und 者 *zhě* formal getrennt in einem Wort auftauchen. Ein Beispiel liefert 面壁者 *miànbìzhě* (*[angesichts] gucken* + *Wand* | jmd.).[7] Bei

[6] Abkürzungen: v = Verb, n = Nomen, bm = Basismorphem, fm = Flexionsmorphem.

[7] 面 *miàn* kann multikategorial ein Nomen, Verb oder Adjektiv sein und verschiedenen Bedeutungen entsprechen. Die Ambiguitätsprozesse werden hier nicht berücksichtigt. Das Wort stammt aus dem Sci-Fi-Roman „Der dunkle Wald" (dem zweiten Band der „Trisolaris-Trilogie") von Liu Cixin und beschreibt einen speziellen Beruf. (Der auserwählte Wandgucker muss aus

der MÜ sollte ein Fuzzy-Matching zwischen „面 *miàn* + 者 *zhě*“ mit „*guck-* + *-er*“ begründet werden, damit das MÜ-System die passende Übersetzung „Wandgucker“ erzeugen kann.

(5) Semantische logische Beschreibungen sind obligatorisch, um zwischen bedeutungsambigen Wörtern/Morphemen auszuwählen. Im Beispiel entspricht 食 *shí* im Deutschen den Verben *essen* und *fressen* und das Wortsuffix 者 *zhě* weist auf eine direkte personale Deutung hin. Anhand der direkten Deutung wird der Wortstamm ‚-ess-‘ ausgewählt (vgl. Lobin 2010: 78ff.).

(6) Modelle der generativen Morphologie sind vonnöten, damit die Wörter in der deutschen Übersetzung anhand des Kontexts entsprechend flektiert werden können.

4 Zusammenfassung und Ausblick

In diesem Artikel wird die DE-ZH-MÜ der Wortbildungen aus Sicht der korpus-basierten MÜ-Methoden analysiert. Zunächst wurden in Kapitel 2 die Wortsegmentationen und inneren Konstituentenstrukturen der beiden Sprachen vorgestellt, die die Voraussetzungen für das interlinguale Alignment der Morpheme darstellen. In Kapitel 3 wurden einige wichtige Punkte der DE-ZH-MÜ von Wortbildungen skizziert: 1) die Verwendung eines ähnlich strukturierten, bekannten Wortes (siehe Absatz 3.1.2), 2) die statistik-basierte mutuale Korrespondenz zwischen Wörtern/Morphemen im fachlichen oder allgemeinen Sprachgebrauch (siehe Absatz 3.1.1 und 3.1.3), 3) die Anwendung und Analyse von Wortbildungsgrammatiken (siehe Absatz 3.2.1), 4) das Fuzzy-Matching zwischen den deutschen Flexions- und chinesischen Morphemen (siehe Absatz 3.1.1 und3.2.2) sowie 5) die Disambiguität nach semantischen Beschreibungen (siehe Absatz 3.2.2).

In der Einleitung wurde erwähnt, dass wegen der hybride verwendeten statistik-basierten SMT- und neuronalen NMT-Technologie Autoübersetzer bei Satzkontexten relativ hohe Übersetzungsqualitäten erreichen. Aber je weniger Kontext mit eingegeben wird, desto häufiger können Übersetzungsfehler auftreten. Insbesondere gibt es keine ideale Lösung bei einzelnen kontextfreien Wortbildungen zwischen dem Deutschen und Chinesischen. Recherche- und Übersetzungsergebnisse sowohl bei Google Translate als auch bei Baidu Fanyi zeigen, dass beide größtenteils über identische Daten verfügen, die zu oft ähnlichen Fehlern bei Wortbildungsübersetzungen führen. Die morphosyntaktischen Analysen und Übersetzungsangebote sind bei Google besser, wenn es um europäische Sprachenpaare geht.

Währenddessen hat Baidu bei der Textverarbeitung des Chinesischen eindeutige Vorzüge. Zielsprachliche Übersetzungen des ‚Westernised Chine-

bestimmten Gründen seine Gedanken vor allen anderen Lebewesen verheimlichen.) Die deutsche Übersetzung von Karin Betz lautet „Wandschauer“.

se‘ von Google und ‚chinglische‘ Ausdrücke von Baidu werden häufig für Wortbildungen zwischen Deutsch und Chinesisch angeboten. Google übersetzt z. B. „das schwarze Kätzchen“ ins Chinesische mit „黑色的小猫“ (*hēisè de xiǎomāo*) viel zu wörtlich: Im allgemeinen Sprachgebrauch würde man eher „小黑猫“ *xiǎo hēi māo* oder „黑猫“ *hēi māo* sagen.[8] Baidu bietet bspw. zu dem Ausgangswort „猫猫“ *māomao* die deutsche Übersetzung „Katze und Katze“ an, was nicht berücksichtigt, dass derlei Verdopplungsstrukturen nur in wenigen Sprachen gebräuchlich sind.[9] Für eine bessere DE-ZH-MÜ reicht die Verbesserung der maschinellen Übersetzung zwischen EN-ZH und EN-DE nicht aus. Zukünftig sind einerseits größere Datenbanken und kontextabhängige sprachliche Verarbeitungen zwischen den drei Sprachen vonnöten und andererseits mehr kontrastiv-linguistische Forschungen der Schriften, Wortbildungen, Syntax, Semantik und des Sprachgebrauchs.

Literaturverzeichnis

Altenberg, B. (1999), Adverbial connectors in English and Swedish: Semantic and lexical correspondences, in: H. Hasselgård und S. Oksefjell (Hrsg.), *Out of Corpora: Studies in Honour of Stig Johansson*, Amsterdam: Rodopi, 249–268.

Bastin, Macheal (2019), *How good is Google translate? The most accurate language pairs*, online: <www.betranslated.com/blog/how-good-is-google-translate/> (Zugang: 08.05.2020).

Berg, Thomas (2012), The cohesiveness of English and German compounds, in: *The Mental Lexicon 7:1*, Amsterdam: John Benjamins Publishing Company, 1–33.

Bußmann, Hadumod (2002), *Lexikon der Sprachwissenschaft*, Stuttgart: Alfred Kröner Verlag.

Féng, Zhìwei 冯志伟 (2001), Quèdìng qīecí dānwèi de mǒuxiē yǔfǎ yīnsù 确定切词单位的某些语法因素 (Some grammatical factors to determine the segmentation elements), in: *Shùyǔ biāozhǔnhuà yǔ xìnxī jìshù* 术语标准化与信息技术 (*Terminology Standardization & Information Technology*), 2/2001, 21–28.

[8] Im Chinesischen müssen die Adjektive – wenn sie gleichzeitig für Größen und Farben monosyllabisch ausgedrückt und zusammen als Determinativ eines Nomens gebraucht werden – die Reihenfolge „zuerst Größe, dann Farbe“ einhalten. Wenn ein Adjektiv bi- oder polysyllabisch ist, muss es nach vorne gesetzt und dahinter optional 的 *de* beigefügt werden. Die von Google angebotene Übersetzung ist zwar nicht falsch, aber unnötig lang und vor allem unpraktisch, wenn der Ausdruck noch erweitert werden soll.

[9] Stand: 10.05.2020; dieselben Fehler sind auch in der englischen Übergangssprache und manchen Satzkontexten aufgetaucht.

Ge, Shili et al. (2018), Clause-Complex Level Error Analysis of English-Chinese Machine Translation, in: Yang Xin-ge et al. (Hrsg.), *Third International Congress on Information and Communication Technology. ICICT 2018*, London/New York: Springer.

Hóu, Mǐn 侯敏 (1999), *Jìsuànjī yǔyánxué yǔ Hànyǔ zìdòng fēnxī* 计算机语言学与汉语自动分析 (*The computational linguistics and the automatic analysis of the Chinese language*), Beijing: Communication University of China Press.

Langer, Stefan (1998), Zur Morphologie und Semantik von Nominalkomposita, in: *Tagungsband KONVENS 98*, Bonn, S. 83–97.

Liú, Chūnhuī 刘春辉 et al. (2009), Jīyú yōuhuà zuìdà pǐpèi yǔ tǒngjì jiéhé de Hànyǔ fēncí fǎngfǎ, 基于优化最大匹配与统计结合的汉语分词方法 (A Chinese segmentation method based on optimization maximum matching and statistics), in: *Yānshān dàxué xuébào* 燕山大学学报 (*Journal of Yanshan University*), 02/33/03, S. 124–129.

Lobin, Henning (2010), *Computerlinguistik und Texttechnologie*, Paderborn: W. Fink.

Popović, Maja, Daniel Stein und Hermann Ney (2006), Statistical Machine Translation of German Compound Words, in: *Advances in Natural Language Processing*, S. 616–624.

Sugisaki, Kyoko und Don Tugger (2018), German Compound Splitting Using the Compound Productivity of Morphemes, in: *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*, Vienna, Austria, September 19-21, 2018.

Trommer, Jochen (2010), Morphologie, in: K.-U. Carstensen et al. (Hrsg.), *Computerlinguistik und Sprachtechnologie. Eine Einführung*, 3. Aufl., Heidelberg: Spektrum Akademischer Verlag, S. 236–263.

Wang, Kai (2019a), *Untersuchungen zur Methodik und Effizienz der tastaturbasierten Eingabeverfahren verschiedener Schriftsysteme der Welt*, Gießen: Gießener Elektronische Bibliothek.

Wang, Kai (2019b), *Die gegenseitigen Entscheidungsfaktoren der Sprachen und der schriftlichen Repräsentation*, Frankfurt am Main: Elektronische Dokumente Universitätsbibliothek, online: <www.publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/50958> (Zugang: 20.08.2019).

Wáng, Xiǎolín 王小林, Yáng, Lín 杨林 und Wáng, Dōng 王东 (2015), Jīyu zhīwǎng de xīn cíyǔ xiāngsìdù suànfǎ yánjiū 基于知网的新词语相似度算法研究 (New Word Similarity Algorithm Research Based on HowNet), in: *Qíngbào Kēxué* 情报科学 (*Information Science*), 2/33, S. 67–71.

Wáng, Xiǎolóng 王晓龙 (2005), *Jìsuànjī zìrán yǔyán chùlǐ*, 计算机自然语言处理 (*The Computer-based Natural Language Processing*), Beijing: Tschinghua University Press.

Wèi, Nǎixīng 卫乃星, Lù, Jūn 陆军 et al. (2014), *Duìbǐ duǎnyǔxue yánjiū: Láizì yǔliàokù de zhèngjù*, 对比短语学探索：来自语料库的证据 (*Phraseology in Contrast: Evidence from English-Chinese Corpora*), Beijing: Foreign language teaching and researching Press.

Xú, Yángchūn 徐阳春 (2008), *Xiàndài Hànyǔ* 现代汉语 (*The modern Chinese Language*), Beijing: Higher Education Press.

Zhōu, Qiáng 周强 und Duàn, Huìmíng 段惠明 (1999), *Xiàndài Hànyǔ yǔliàokù jiāgōng zhōng de qiēcí yǔ cíxìng zìdòng biāozhù*, 现代汉语语料库加工中的切词与词性自动标注处理 (*The Segmentation and POS-Tagging in the Processing of the modern Chinese Corpus*), Beijing: Běijīng dàxué jìsuànyǔyánxué yánjiūsuǒ 北京大学计算语言学研究所 (Institute of Computational Linguistics [ICL]).

Analysis of Mutual Machine Translation between German and Chinese Word Formations and Improvement Possibilities

Abstract

When it comes to writing, grammar and culture, the German and Chinese languages differ vastly from one another, and incorrect or unintelligible translations are a commonplace due to a lack of bilingual corpora. Unregistered and ad-hoc word formations make up a large portion of translation errors.

This paper aims to answer the question why and how artificial intelligence (AI) makes translation errors when text translating between German and Chinese. Why are word segmentations and divisions necessary and how do they work in German and Chinese corpora? How can AI use annotated corpora, registered words and linguistic rules to produce higher quality translations of unregistered word formations and ad-hoc structures between the two languages?

Keywords: DE-ZH-MT, machine translation, alignment, word formation, segmentation

论中德文合成词在机械翻译中的问题与优化方法

摘要

由于中德文在文字、语法与文化上的巨大差异，且中德双语平行语料库在机械翻译中的应用不足，以致汉德机械互译中时常出现错译或翻译不明的情况，其中，未登录合成词及临时合成词的翻译错误占据较大比重。

本篇将重点探究如下问题：人工智能如何在中德文合成词的翻译上“犯错”？为何词语切分和合成词内部划分是不可或缺的步骤，中德文语料库的词语划分又如何进行？在熟语料库（即已经过分词和词性标注处理过的语料库）、已收录词、语言规则的基础上，人工智能如何能在互译中输出质量较高的译文？

关键词：中德机械互译、自动翻译、词语对其、合成词、词语切分

Manuskript eingereicht am 25.02.2020; akzeptiert am 25.05.2020

Anhang

Abkürzungen

DE-ZH-MÜ: maschinelle Übersetzung zwischen Deutsch und Chinesisch
EN-ZH-MÜ: maschinelle Übersetzung zwischen Englisch und Chinesisch

Terminologische Erklärung

Wortbildung: „Bildung neuer komplexer Wörter auf der Basis vorhandener sprachlicher Mittel[.]“ Darunter machen Derivation und Komposition den größten Teil aus. (Bußmann 2002: 751)
Unregistrierte Wörter: Da jedes Wörterbuch einen begrenzten Umfang hat, gibt es immer Wörter, die nicht von einem Wörterbuch aufgenommen werden. Solche in Wörterbüchern nicht eingetragene Wörter sind unregistrierte Wörter.
Segmentation: Segmentation bedeutet die Zerlegung komplexer Einheiten in kleinere Einheiten. In dieser Arbeit werden zwei Arten von Segmentation berücksichtigt: die Wortsegmentation (die Zerlegung der Texte in Worteinheiten) und die Segmentation der Worteinheit in Morpheme.