22W 520.005 Transkationsrelevantes Informationsmanagement

Andreas Hofer

January 25, 2023

1 Vorstellung/Translationsrelevante Technologien

1.1 Organisatorisches

Das gesamte Material, also Folien, werden stets hochgeladen, ist jedoch für die Prüfung nicht ausreichend und es werden wahrscheinlich auch Notizen vonnöten sein um die Prüfung zu bestehen.
Die Vorlesung ist Vortragsbasiert, also gibt es keine Bewertung während des Kurses, jedoch werden im Jänner die Vorlesungen online sein, wodurch man die Werkzeuge gleich am PC testen kann.
Der Text von Sharon O'Brien über Fehler in der Maschinentranslation ist 16 Seiten lang und ist bis zum 14.12 zu lesen. Dieser ist auch relevant für die Prüfung. Die Prüfung selbst ist in Präsenz und Single-Choice. Er wird Praxisorientiert sein.
Es gibt insgesamt 6 Einheiten, wobei 3 vor Weihnachten in Präsenz stattfinden werden. Nach Weihnachten gibt es drei weitere Einheiten, welche online stattfinden werden.
Der Text ist einem längeren Buch entnommen, welches Open-Access ist weshalb man überall frei darauf zugreifen kann.

1.2 Was für Übersetzungstechnologien gibt es?

Heutzutage gibt es einige Technologien mit denen Übersetzt werden kann, oder die einem bei der Übersetzung helfen. Einige Werkzeuge sind DeepL oder Google Translate, welche automatisch übersetzen. Werkezuge die bei der Übersetzung helfen sind Wörterbücher wie Linguee, Reverso aber auch reine Wörterbücher wie Pons oder der Duden.

1.3 Wie teilt man Translationstechnologie ein?

Die Reichweite von Übersetzungswerkzeugen ist außerordentlich groß. Werkzeuge zur Übersetzung sind zwar die beliebtesten, wo es mehrere dutzend unterschiedliche Anbieter gibt, jedoch gibt es auch eine größere Zahl an Anbietern für unter anderem Maschinenlerntechnologie. Übersetzungstechnologien können in zwei Kategorien unterteilt werden:

- Computergestützte Technologien (CAT)
  - Diese kann man weiter einteilen:
    - Nach Grad der Automatisierung: Sind Menschen mehr oder weniger im Prozess des Übersetzens involviert? Es reicht von Vollautomtischen Übersetzern bis zur traditionellen Humantranslation, wo man keine Hilfsmittel verwendet. In der Mitte dieser liegen die *Computer-assisted translation technologies* Machine-aided human translations und human-aided machine translations, abhängig davon werlcher Anteil größer ist. Im Feld des literarischen Übersetzens wird immer noch größtenteils ohne Hilfsmittel übersetzt, da die Übersetzung sehr vom Kontext abhängig ist. Man kann diese Werte als Tabelle darstellen, indem zwischen *Human Translation*, *Computer-aided Translation*, *Machine Translations* unterschieden wird. Den größten Teil macht hierbei das CAT aus, da es einen Übersetzer bei der Tätigkeit unterstützt, was meist die beste Kombination aus Zeit und Effizienz ist.
    - Nach Verwendung im Übersetzungsprozess: *Before-Translation*, *During-Translation*, *After-Translation*
    - Nach der Beziehung zum Übersetzungsprozess: Wie Allgemein oder Spezifisch ist die Technologie? Ist es Allgemein wie z.B. Microsoft Word oder ist es spezifischer wie PONS oder Linguee. An unterster Stelle liegen Programme spezifisch für professionelle Übersetzer wie Trados oder Antconc
    - Nach der Dimension der Translation: Werden die Werkzeuge für die Lehre, professionelle Übersetzer oder im Bereich der Forschung des Übersetzens verwendet?

1.4 Geschichte der Translationstechnologie

Maschinelle Übersetzung besteht theoretisch schon seit den 30er Jahren als die Ersten PCs entwickelt wurden. Die Idee des automatischen Übersetzens ist jedoch bedeutend älter wobei Leibniz und Descartes bereits Überlegungen darüber gemacht hatten. **George Artsrouni** patentierte 1930 das *Mechanical Brain*. In diesem ersten Versuch wurden spezifische Phrasen abgespeichert, wodurch es nur einen limitierten Anwendungsbereich hatte, jedoch die Basis für spätere Forschung bildete. Gleichzeitig überlegte der Russe **Petr Petrovich Smirnov-Troyanskij** über eine Übersetzungsmaschine, welche Wörter konjugieren kann, konnte es jedoch nie realisieren.

1.4.1 Das Weaver Memorandum

Warren Weaver, bekannt vom Warren Weaver Kommunikationsmodell, überlegte auch was eine Übersetzung ausmacht. Er hatte mehrere Überlegungen:

1. Eine Übersetzung muss Kontextbasiert sein, also muss es eine Bank für Geld von einer Bank zum sitzen unterscheiden können
2. Sie muss logische Komponenten der Sprache abbilden können
3. Da zu dem Zeitpunkt noch Kodewechsel populär war, dachte er auch über universelle Bedeutungselemente nach, welche in jeder Sprache das selbe bedeuten. Wenn man diese Elemente nun findet könnte eine Maschine jede Sprache in jede Sprache übersetzen. Das hat sich später als fehlerhaft herausgestellt, wodurch diese Überlegeung hinfällig wurde
4. Es gibt Sprachuniversalia. Dass es Konzepte gibt, welche in jeder Sprache vorhanden sind. Wenn man also die Universalia definiert sollte man diese Worte einfacher übersetzen können.

1.4.2 Erste Schritte

In den 40er Jahren wurden die ersten Überlegungen von maschineller Übersetzung getätigt, wie eben das Weaver Memorandum. Zu diesem Zeit wurde auch Übersetzung als Kodewechsel gesehen.
1954 wurde das erste System zur Übersetzung vom Englischen zum Russischen vorgestellt, größtenteils um in der Zeit des Kalten Krieges schneller an Informationen zu kommen. Zu diesem Zeitpunkt war man der Ansicht, dass es bald zu einem Durchbruch in der maschinellen Übersetzung kommen würde. Nachdem dies nicht so schnell passierte, wurde das *ALPAC*(Automatic Language Processing Advisory Committee) gegründet. Dieses Kommittee kam zu dem Schluss, dass vollautomatische Übersetzung nicht so schnell erreichbar ist, und man stattdessen in Menschliche Hilfswerkzeuge zur Übersetzung investieren sollte. Dadurch wurde ein Großteil der Forschungsgelder in den USA von maschineller Übersetzung abgezogen, welches zu diesem Zeitpunkt Vorreiter in der Technologie war. Andere Länder wie Japan führten diese Forschung jedoch weiter.
In den USA wurde aber sehr wohl die Translation Memory entwickelt, welche als Glossar für Übersetzer dienen sollte.
In Europa gab es auch Forschung für Terminologiedatenbanken wie EURODICAUTOM, welcher Vorgänger des IATE war. Da diese Forschung jedoch in alle Sprachen der Mitgliedsstaaten übersetzt werden musste, war der Prozess etwas verlangsamt. Eine weitere für Rechtstexte von Kanda ist TERMIUM.
Das IATE ist heute noch öffentlich zugänglich und kann verwendet werden um spezifische Fachtermini in allen Sprachen der EU nachzusehen. So kann man nachsehen, was zum Beispiel die GDPR in verschiedenen Sprachen heißt, wie gut dieser bewertet wurde und woher er stammt.
In den 80er Jahren tauchten komerzielle Übersetzungstools auf und TRADOS, welches seit 1984 besteht, ist heute noch ein populäres Werkzeug. Zwar hat es vor 40 Jahren sehr anders ausgesehen, doch ist die Kernfunktionalität noch immer die selbe. So versucht das Programm bereits getätigte Übersetzungen zu speichern und den Benutzer darauf hinzuweisen wenn die momentane Übersetzung mit einer früheren ähnlich ist. Es weiß auch, wenn nur ein Teil der Übersetzung ähnlich ist.

Ab den 1990er Jahren gab es immer mehr Anbieter solcher Technologie, wobei diese Tools auch mehr Funktionalitäten erhielten. Zwar ist die Speicherung von Übersetzungen die Grundfunktionalität, doch kamen mit der Zeit immer mehr ünterstützende Funktionen hinzu. Eines davon ist die Alignierung, mit welchem man externe Texte in das Tool einpflegen kann um diese auch innerhalb des Programms verwenden kann.
Mit der Zeit wurden diese Tools immer benutzerfreundlicher und es wurde versucht die Tools an den Workflow der Übersetzer anzupassen, anstatt zu erwarten, dass die Menschen sich der Maschine anpassen. Auch wurden Cloud-Funktionen und Integration in andere Programme realisiert. Gleichzeitig wurden offene Standards entwickelt um Terminologiedatenbanken mit anderen Leuten als Datei zu Teilen. Maschinelle Übersetzung erfuhr durch den ALPAC-report einen Dämpfer in den USA. In Japan wurde diese Technologie jedoch weiterentwickelt, und in den 1990er Jahren wurden die ersten statistischen Übersetzungstechnologien veröffentlicht. Google stieg erst 2007 auf statistische Übersetzungstechnologie um und 2016 auf neuronale Übersetzungstechnologien. Dass 2016 neuronale Übersetzer besser geworden waren, wurde durch das maschinelle Qualitätssicherungsverfahren BLEU

2 Information, Wissen, Technologie und Translation

Ein jeder Übersetzungsauftrag erweitert das Wissen eines jeden Übersetzers, da man sich stets sprachlich und kulturell weiterbilden muss. Da man bei der Übersetzung nicht nur Wörter, sondern den Sinn übersetzt, muss man den Text zuerst verstehen, bevor er übersetzt werden kann.

2.1 Wissensmanagement

Wissensmanagement ist die Ansammlung von Methoden um effektiv übersetzen zu können. Hierbei gibt es einige Grundbegriffe:

- Daten
  - Zeichen aus einem Datenvorrat, welche durch Syntax kombiniert werden
- Information
  - Die Syntax um diese Zeichen miteinander zu kombinieren.
- Wissen
  - Der Kontext, welcher die Zeichen und die Syntax im richtigen Weg anwendet
  - Geteilt in wiederum zwei Kategorien:
    - Explizites Wissen: Eindeutig erlerntes Wissen, wie es in einem Lehrbuch gefunden werden kann
    - Implizites Wissen: Nicht offensichtliches Wissen welches indirekt weitergegeben wird.

Wissensmanagement wird auf unterschiedlichen Ebenen angewandt. Wissensmanagement innerhalb einer Organisation beispielsweise wird größtmöglich effektiv untereinander geteilt um die Oragnisation weiterzubringen? Persönliches Wissensmanagement hingegen ist das Wissen, welches eine Person benötigt um erfolgreich zu sein. Zum Beispiel wenn man selbstständig ist.
Es wird stets danach gestrebt, Wissen zu personalisieren. Implizites Wissen ist ein Prozess, welcher es vermittelt, jedoch findet man dieses oft nicht in einem Buch in einer Bilbiothek.
Es gibt verschieden Arten von Wissen:

- Sprachliche, linguistische und übersetzerische Kenntnisse
  - Kann in kodifiziertes
    - Kann in Lehrbüchern gefunden werden und stellt das Basiswissen dar, welches für Übersetzung relevant ist.

  - und nicht-kodifiziertes Wissen eingeteilt werden:
    * Kontextverständnis oder Bedürfnisse der Zielleserschaft. Diese können nicht explizit mit einem Buch erlernt werden und müssen mit der Zeit erarbeitet werden.

- Landes- und Kulturwissen
  - Kodifiziert:
    * Variationen zwischen Ländern wie British und American English
  - Nicht Kodifizierbar:
    * Moralische Einstellungen beziehungsweise das politische Umfeld. Übersetzungsverbände und kulturelle Veranstaltungen könenn verwendet werden um sich dieses Wissen anzueignen.

- Allgemeines Sachwissen
  - Kodifiziert:
    * Industrienormen zwischen Ländern oder spezifisches Fachwissen. Diese können in Ontologien dargestellt werden. Diese zeigen Verhältnisse zwischen Gegenstandsbereiche und zeigen aktive und passive Prozesse.
  - Nicht kodifiziert:
    * Hausverstand und Erfahrung, welches durch bspw. Plattformen

- Kunden und firemnspezifisches Wissen

In diesen Studien wird festgelegt, welche Kompetenzen Übersetzer benötigen um übersetzen zu können. In Europa besteht der EMT Kompetenzrahmen, welcher in Zusammenarbeit mit Übersetzungsorganisationen entstanden ist.
Studien die sich diesem Kompetenzrahmen verpflichten, lehren anhand dieses Rahmens, welcher darüber schreibt, dass ein Übersetzer bspw. wissen sollte wie maschinelles Übersetzen funktioniert.

2.2 Datenbestände

2.2.1 Lexikalische Datenbestände

Lexikalische Datenbestände basieren auf der Lexikographie, welche sich mit Wörterbüchern beschäftigt. Computergestützte Lexikographie besteht schon seit den 40er Jahren, haben sich in den letzten 80 Jahren jedoch bedeutend weiterentwickelt. So haben sich einige Standards with TEI P5 durchgesetzt, mit welchem man Wörterbucheinträge standardisiert formatieren kann.
Bei Elektronischen Wörterbüchern kann man zwischen einigen Sachen unterscheiden:
Ob mit den Wörterbüchern Menschen oder Computer angesprochen werden sollen. Zusätzlich unterscheidet man zwischen Retrogradisierten oder neu-digitalisierten Wörterbüchern. Ob das Wörterbuch digitalisiert wurde oder komplett neu als digitales Wörterbuch konzipiert wurde.
Besteht das Wörterbuch aus freiwilligen Nutzern oder wird diese redaktionell betreut.
Im Gegensatz zu Wörterbüchern gibt es auch Terminologische Resourcen, welche oft riesige Korpora darstellen und auch oft öffentlich einsehbar sind. Innerhalb der Terminologie unterscheidet man zwischen Gemein- und Fachsprache, welche zwischen Termini und Wörter geht. Wenn eine Fachsprache eine spezielle Definition für ein Gemeinsprachliches Wort hat, ist das ein Termini.
Man muss auch zwischen Begriff, Benennung und Gegenstand unterschieden. Ein Gegenstand ist ein beliebiger Ausschnitt aus der wahrnehmbaren Welt: Ein Haus kann gleich ein Gegenstand sein wie das Konzept der Freundschaft. Diese Gegenstände werden durch Begriffe definiert. Diese Begriffe existieren nur in den Gedanken der betroffenen Personen. Erst wenn die Benennung ins Spiel kommt wird mit dem Begriff ein spezifisches Wort verbunden. Hier werden Terminologie und Lexikographie unterschieden. Bei der Terminologie steht die Begriff an erster Stelle, während in der Lexicographie die Benennung wichtiger ist.
Also beginnt die Lexikographie mit der Bezeichnung und leitet von dort die Benennung ab, während es bei der Terminologie umgekehrt ist.

Terminologiedatenbanken haben auch ihren eigenen Standard, sodass das Termbase Exchange Format (TBX) universell verwendbar ist.
Mittels diesem Standard kodieren Unternehmen ihre Datenbanken manchmal und machen diese öffentlcih zugänglich.

2.3 Translation Memory Tools

Während CAT tools und Translation Memory Tools oft synonym verwendet werden, ist CAT der Überbegriff und wenn Leute von CAT Tools sprechen, meinen sie oft TMTs. Diese Tools funktionieren so, dass man bereits getätigte Übersetzungen abspeichern kann und das Programm, wenn ein gleicher oder zumindest ähnlicher Satz übersetzt werden muss, diesen einem anzeigt. Solche Datensätze bestehen stets aus Metadaten und einem Übersetzungspaar.
Die bekanntesten Toolsfür diesen Zweck sind *Déja Vú*, *MemoQ* oder *Memsource*.
In der GUI wird das Interface in zwei Segmente geteilt. Auf einer Seite hat man den Originaltext und ein EIngabefeld zur Übersetzung, während der andere die Datenbank durchsucht und eventuelle Treffer anzeigt.
Man nennt diese Tools zwar immer noch Translation Memory Tools, doch haben solche Programme heutzutage viel weitere Möglichkeiten.
Solche Translationtools bestehen bereits seit den 80er Jahren mit TRADOS, welches auch heute noch Makrtführer ist.
TMTs bestehen aus verschiedenen Komponenten:

- Übersetzungsspeicher
  - Können lokal oder in der Cloud gespeichert werden. So kann jeder die aktualisierte Version abrufen.
- Terminologiedatenbank
- Alignierungsprogramm
  - Wenn man bereits einen großen Corpus hat, dieser aber nicht Teil des Programms ist, gibt es sogenannte Alignierung um solche Korpora zu importieren. Traditionell wird der Zeiltext zuerst Segmentiert und dann diese Segmente in die Datenbank importiert.
- Maschinelle Übersetzung

Die TMTs speichert den Korpus in kleineren Segmente, da das die Abspeicherung vereinfacht.
Wenn das TMT die Datenbank durchsucht unterscheidet man zwischen 100% Matches und fuzzy matches. Bei einem 100% Match wurde in der Vergangenheit der genau gleiche Satz übersetzte, während es bei einem fuzzy match nicht unbedingt der Fall sein muss. Man kann es individuell einstellen, wann man einen fuzzy match haben will.
Es gibt mehrere Wege um so ein TMT verwenden zu können. So gibt es ein Add-On zu Microsoft Word namens WordFast, aber auch alleinstehende tools wie die vorhin genannten MemoQ.
Diese Tools müssen nicht unbedingt auf einem PC installiert werden und man kann gewisse Tools, welche über das Intranet einer Organisation oder Web/Cloud basiert ist.

2.4 Anwendungsbeispiele

TMTs sind sehr nützlich, wenn man immer wieder mit dem selben Kunden zu tun hat oder oft einen ähnlichen Texttypen aufweisen, da man dann direkt auf die Datenbank zugreifen kann.

3 Berechenbarkeit des Übersetzens

3.1 Relevanz der MÜ für die TLW

2018 wurde ein paper veröffentlicht, welches behauptete Parität zwischen Nachrichtenübersetzung erreicht zu haben. Es stellte sich heraus, dass die Forscher fragwürdige Praktiken angewandt hatten damit es so aussieht, als ob die MÜ besser

wäre, als sie wirklich war. Während dieses Beispiel nicht zu der gewünschten Lösung führte, zeigt es, dass man sehr wohl schon darüber nachdenken kann, ob MÜ gleich gut ist als menschliche Übersetzungen.
Da maschinelle Übersetzung größtenteils auf bereits getätigten menschlichen Übersetzungen basiert, entsteht eventuell ein ethischer Konflikt da es im Endeffekt dazu führt, dass die Arbeit auf Maschinen ausgelagert wird.

3.2 Maschinelle Übersetzung?

MÜ hatte lange Zeit eine andere Bedeutung innerhalb der TLW, da Übersetzung als mechanischer Prozess (Kodewechsel) gesehen wurde. Erst mit dem Cultural Turn änderte sich dies.
Es gibt mehrere Zugänge innerhalb der MÜ:

- Technikdeterminismus: Die Existenz der Technologie gibt den Anstoß zum Finden von Anwendungsmöglichkeiten.
- Sozialkonstruktivismus: Dass die Bedeutung der MÜ nicht in der Maschine selbst, sondern im Kontext der sozialen Gruppen geschaffen wird. So erhält diese ihre Aufgabe und Bedeutung erst im sozialen Kontext.

Translationskonzepte und Übersetzungsbegriffe existieren nicht nur innerhalb der TLW. So haben Kulturwissenschaften und die Computerlinguistik ein Verständnis dessen.
Die Kulturwissenschaften verstehen Translationskonzepte als jeglichen Übergang von kulturellen Dingen als Übersetzung und so eine sehr breite Definition dessen.
Die Computerlinguistik definiert diese hingegen sehr eng.
In der maschinellen Übersetzung kann man von beiden Seiten ausgehen, sowie davon ausgehen, ob die Technik den Menschen beeinflusst oder umgekehrt.
Innerhalb einer Studie über maschinelle Translation definierten unterschiedliche Befragte innerhalb der Diszplin der maschinellen Übersetzung diese mehr oder weniger eng. Eine(r) sagte, dass die Übertragung von Information in eine Sprache in eine andere Übersetzung ist. Eine(r) andere(r) sagte jedoch, dass die Vermittlung zwischen zwei Kulturen Übersetzen ist.
Ebenfalls wurden die Unterschiede zwischen maschineller und humaner Übersetzung erfragt. Manche sagten, dass maschinelle Übersetzung nicht von Verständnis rührt, sondern von statistischer Auswertung humanübersetzte Texte. Die Maschine wendet Statistik an und versteht nicht den Kontext dahinter. So wurde zwischen Textverständnis und Weltmodell bei Humanübersetzung und Statistischer Auswertung bei maschineller Übersetzung unterschieden.
Letztlich stand die Frage, ob Technologie den Rahmen vorgibt in welchem es verwendet wird, oder der soziale Kontext diesen vorgibt. Ein Befragter sagte, dass neuronale Netzwerke so vorherrschend sind, dass Wissenschaftler nicht darüber nachdenken wie man ein Problem löst und dann die Lösung erarbeitet, sondern stattdessen über die Awendungsmöglichkeiten.
Auf der anderen Seite war die erste große Anwendung der MÜ aufgrund des Kalten Krieges sehr erwünscht, da die USA eine einfache Möglichkeit wollten um Russisch zu verstehen. Dies ist ein Beispiel von Sozialkonstruktivismus. Ein weiteres ist Forschung an Universitäten, welche Technologie erforschen, nicht weil sie es anwenden wollen oder können, sondern weil es sie interessiert.

3.3 Post-Editing

Post-Editing ist die Nachbearbeitung von Maschinell-übersetzten Texten zur Qualitätskontrolle. Die Technologie selbst ist gerade in einem starken Wandel, weshalb die Praktiken noch nicht gänzlich festgelegt sind.
Post-Editing kann in ein- und mehrsprachliche unterteilt werden:

- Mehrsprachlich: Der übersetzte sowie originaltext liegen vor und der sprachliche und kontextuelle Inhalt kann überprüft werden.
- Einsprachlich: Es liegt nur der übersetzte Text vor und man kann diesen auch nur anhand sprachlicher Richtlinien anpassen.

Ebenfalls existiert light und full post-editing, welches wiederum viel oder wenig an diesem verändert. Bei light post-editing verändert man nur die sprachlichen eigenschaften sodass, ein verständlicher Text entsteht. Bei full post-editing hingegen wird darauf abgezielt das äquivalent eines vollständig professionell übersetzen Text zu erhalten.
Light post-editing versucht meist "good enough" zu erzielen. Good enough bedeutet, dass die Grammatik, die Rechtsschreibung und andere textuelle Sachen richtig sind und dieser lediglich sprachlich verständlich ist.
Full post-editing hingegen versucht "comprehensible" zu sein, sodass dieser zusätzlcih zu good-enough auch inhaltlich verständlich und inhaltlich vergleichbar mit dem AT ist. So muss der Text auch syntaktisch verständlich sein. Jedoch gibt es keine Erwartung, dass der Text einen ähnliche Qualität wie die eines Muttersprachlers aufweist.
Was bei beiden der Fall sein muss ist, dass keine Bedeutung verloren geht oder hinzugefügt wird und möglichst viel des MÜ übersetzten Textes verwendet wird. Da post-editing das Ziel hat den menschlichen Aufwand zu veringern, soll man sich so viel wie möglich an diesem orientieren.
Bei Post-Editing bestehen zwei Typologien:

- Adequacy
  - Fehler in der Bedeutung von Wörtern oder ein translation-shift
- Acceptability
  - Bezieht sich auf die sprachliche Umsetzung, sodass der Stil und die Flüssigkeit des Textes stimmt.

4 Maschinelle Übersetzung

Maschinelle Übersetzung teilt sich in drei große Teile:

- Regelbasiert
- Statistisch
- Neuronal

Die ersten Systeme waren regelbasierte Systeme. Die allerersten Systeme waren überhaupt nur ein Text, welcher mit einem Wörterbuch verglichen wird, wodurch die Wörter übersetzt werden. Spätere Systeme stellten sicher, dass diese ebenfalls der Zeilsprachlichen Grammatik entspricht. Ein weiteres theoretisches Modell, welches jedoch in der Praxis nie wirklich umgesetzt wurde, ist das Interlingua-Modell, also dass eine einheitliche Zwischensprache als Medium verwendet wird. So wird der Ausgangstext zuerst in eine abstrakte Repräsentation gebracht, welcher ein Zahlencode, aber auch eine künstliche Sprache wie z.B. Esperanto sein kann und später in die Zielsprache übersetzt wird. Der Vorteil wäre hier, dass man ein gemeinsames Modell der Übersetzung hat wodurch über die Mediumssprache jede Sprache in jede andere übersetzt werden kann.
Ein Modell zur Darstellung von Regelbasierten Systemen ist die Vauqouis Pyramide, in welcher der Verlauf von einem Transfer zu einer direkten Übersetzung dargestellt wird. So wird es mehr zum Transfer, je mehr der zu übersetzende Text analysiert wird und das endgültige Ziel dieser Analyse wäre die Erstellung eines Interlinguamodells.
Regelbasierte Systeme funktionierten leider nicht außergewöhnlich gut, da man einen sehr großen Aufwand hatte, jedoch keine besonders guten oder akkuraten Systeme bekam.
Ein alternativer Weg um diese Regeln zu erlangen ist dem System einen Korpus zu geben wodurch sich dieses dann seine Regeln selbst ableitet. Dieser Korpus muss aligniert werden, also die übersetzten Teile den originalsprachlichen Teilen angepasst werden. Zusätzlich zu diesem zweisprachigen Korpus benötigt das System jedoch auch noch ein einsprachiges Korpus damit es weiß, wie die Sprachfolge in der Zeilsprache ist. Also benötigt ein Statistikbasiertes System jeweils einen Übersetzungs und einen Referenzkorpus.
Statistische Systeme basieren auf zwei Systemen:

- Wortbasierte Systeme
  - Daten werden auf Ebene der Wörter analysiert, wodurch jedem Wort in der Quellsprache ein Wort in der Zielsprache entspricht. Das funktioniert jedoch nur, wenn es genau ein Wort für ein anderes Wort gibt, da Synonyme nicht akkurat dargestellt werden können.
- Phrasenbasierte Systeme
  - Indem Phrasen (Welche eine gewisse Anzahl an Wörtern ist) analysiert werden, kann das System den Kontext von gewissen Wörtern analysieren. Das System kann so Disambiguierungsentscheidungen treffen und zwischen Wörtern wie Bank (Der Institution) und Bank (Der Sitzgelegenheit) unterscheiden.

Der Vorteil von diesen Systemen ist, dass man anstatt der Transferregeln selbst generieren zu müssen, nur die Korpora bereitstellen muss. So hat man bedeutend weniger Aufwand, bei oft besseren Ergebnissen.
Nachteile sind, dass es nur für Sprachen möglich ist, für welche genügend Texte verfügbar sind. Kleinere und nicht umfangreich erforschte Sprachen haben oft nicht genug Textmaterial zur Verfügung um ein Regelwerk erstellen zu können. Ebenfalls funktioniert es nicht so gut bei stark flektierten Sprachen, wobei man jedoch mit Grammatikmodulen diesen Nachteil ausgleichen kann.
IBM begann 1980 mit der Entwicklung von statistischen Maschinen, und diese waren lange die akkuratesten Systeme. Erst in 2015 stiegen die meisten Übersetzer auf neuronale Systeme um.
Neuronale Systeme kann man als Weiterentwicklung von statistischen Systemen sehen, da diese auch große Mengen an Daten benötigen, aber statt Statistik versucht Zusammenhänge wie in einem neuronalen Netzwerk auszuwerten. Diese Systeme sind heutzutage am akkuratesten, da die Bedeutungsähnlichkeit von Wörtern genauer erfasst werden kann.
Neuronale Systeme verwenden Bedeutungsräume um Wörtern Bedeutungen zuzuweisen. Ein Bedeutungsraum kann zum Beispiel sein, dass röt und grün adejtkive sind, da diese oft mit Substantiven verwendet werden. Gleichzeitig weiß das System wo und in welcher Weise gewisse Wörter in welchem Kontext verwendet werden. Diese Bedeutungsräume werden dann für beide Sprachen zugeordnet und dann aneinander angeglichen. In der Mathematik spricht man hier von einer *Zuordnungsfunktion*. Diese Räume werden schließlich in Neuronalen Netzwerken verkettet, welches man trainieren muss. Dieses Training gibt die Aufgabe und die Lösung gleichzeitig an, wodurch das System lernt welche Aufgaben welche Lösungen verlangen. Dadurch werden gewisse Pfade innerhalb des Netzwerkes öfters beansprucht was diese stärkt. Das passiert millionen von Malen, wodurch das System stets besser wird.
Neuronale Systeme sind sehr gut in Sprachen, in welcher Wörter sehr voneinander Abhängen, wie zum Beispiel im Deutschen. Nachteile sind das begrenzte Vokabular der Systeme, da es oft verwendete Wörter selten verwendeten Wörtern vorzieht. Gleichzeitig sind seltenere Wörter oft sehr schwer zu übersetzen, da diese einfach nicht so oft vorkommen. Auch ist das Training von neuronalen Netzwerken extrem rechenintensiv, wodurch auch oft nur riesige Technologiekonzerne die resourcen haben diese durchzuführen. Zusätzlich haben sie, gleich wie beis statistischen Systemen, Probleme Sprachen mit kleinen verfügbaren Korpora zu übersetzen.

XML-Dokumente haben stets den selben Aufbau. Sie basieren auf Tags, welche die Information umgibt. So kann man *<speaker>* und *</speaker>* um einen Namen schreiben, wodurch das System diesen dann richtig klassifizieren kann. Tags bestehen stets aus einem Ausdruck, welcher mit < und > umgeben ist. Der beendende Tag hat zusätzlich noch ein / um das zu kennzeichnen.

Korpustypologie beschreibt den Typ von Korpus, um den genau benötigten Zweck zu erfüllen.
So gibt es verschiedene Kropustypen:

- Lernerkorpus
  - Korpus um eine Sprache zu lernen mit annotationen
- Parallelkorpus
  - Korpus aus zwei oder mehr Sprachen, um diese vergleichen zu können

- Translation Memory Daten
  - Verbindungen aus früher getätigten Übersetzungen

Der Unterschied zwischen Parallelkorpora und Translation Memory Daten ist, dass der Korpus stets im Volltext verfügbar ist. Selbst wenn Teile des Korpus mehrmals vorkommen, werden diese immer wieder gespeichert. In TM stattdessen wird, nachdem man einen Teil ein Mal verwendet hat, dieses immer nur ein Mal abgespeichert.