The forms of words within a sentence are systematic, words in this example pretty much translate one-for-one. But we have to make sure:

- Take the simple Spanish sentence and its English translation:

  - Words are grouped into meaningful units. (cf., our word order can differ from language to language.
  - Things that may be a life or death situation:
    - moving texts from one (human) language (e.g., search engines)
    - diplomatic negotiations
    - court proceedings
    - . . .
  - Words have to be translated.
  - Things that require subtle knowledge of the world:
    - obtaining information from sources in multiple languages (e.g., search engines)
    - pharmaceutical business
    - automatically translating frantic 911 calls for a caller who speaks only Spanish

What is Machine Translation?

- Translation is the process of:
  - moving texts from one language (source language) to another (target language),
  - in a way that preserves meaning.
  - Machine translation (MT) automates (part of) the process:
    - Fully automatic translation
    - Computer-aided (human) translation

What is MT good for?

- When you need the gist of something and there are no human translators around:
  - translating e-mails & webpages
  - obtaining information from sources in multiple languages (e.g., search engines)
  - If you have a limited vocabulary and a small range of sentence types:
    - translating weather reports
    - translating technical manuals
    - translating terms in scientific meetings
    - determining if certain words or ideas appear in suspected terrorist documents → help pin down which documents need to be looked at closely
  - If you want your human translators to focus on interesting/difficult sentences while avoiding lookup of unknown words and translation of mundane sentences.

What is MT not good for?

- Things that require subtle knowledge of the world:
  - translating Shakespeare into Navajo
  - diplomatic negotiations
  - court proceedings
  - . . .
  - Things that may be a life or death situation:
    - pharmaceutical business
    - automatically translating frantic 911 calls for a caller who speaks only Spanish

Example translations

- The simple case:
  - Yo hablo español.
    - I speak Spanish.
    - Words in this example pretty much translate one-for-one.
    - But we have to make sure hablo matches with Y o, i.e., that the subject agrees with the form of the verb.

- The order and number of words can differ:
  2a. Tu hablas español?
    - You speak Spanish?
  2b. Hablas español?
    - Speak Spanish?

Example translations

- A slightly more complex case:
  - Words have to be translated. → dictionaries
  - Words are grouped into meaningful units. (cf., our discussion of syntax for grammar checkers)
  - Word order can differ from language to language.
  - The forms of words within a sentence are systematic, e.g., verbs have to be conjugated, etc.
Different approaches to MT

- Transformer systems
- Systems based on linguistic knowledge
  - Direct transfer systems
  - Interlinguas
- Machine learning approaches

Most of these use dictionaries in one form or another, so we will start by looking at dictionaries.

An MT dictionary is differs from a "paper" dictionary:

- must be computer usable (electronic form, indexed)
- contain the inherent properties (meaning) of a word
- need to be able to handle various word inflections
  have is the dictionary entry, but we want the entry to specify how to conjugate this verb.

Dictionaries

An example (cont.)

3. Using the dictionary, find the target language words
   (6) Drehen Sie [den Knopf] [eine Position] zurück.

4. Using the source-to-target rules, reorder, combine, eliminate, or add target language words, e.g.,
   - 'back' goes with 'turn'; reorder 'back' after the knob'
   - because 'Drehen ... zurück' is a command, in English it
     is expressed without 'you'.

⇒ End result: *Turn the knob back one position.*

Transformers: Less than meets the eye

- By their very nature, transformer systems are non-reversible because they lack a target language grammar.
  If we have a German to English translation system, for example, we are incapable of translating from English to German.
- However, as these systems do not require sophisticated knowledge of the target language, they are usually very robust = they will return a result for nearly any input sentence.
Linguistic knowledge-based systems

- Linguistic knowledge-based systems include knowledge of both the source and the target languages.
- We will look at direct transfer systems and then the more specific instance of interlinguas.
  - Direct transfer systems
  - Interlinguas

Direct transfer systems

- A direct transfer systems consists of:
  - A source language grammar
  - A target language grammar
  - Rules relating source language underlying representation to target language underlying representation

Steps in a transfer system

1. source language grammar analyzes the input and puts it into an underlying representation (UR).
   Der Tisch gefallen Paul → Der Tisch gefallen Paul (source UR)
2. The transfer component relates this source language UR (German UR) to a target language UR (English UR).
   German UR: X gefallen Y → Eng(Y) like Eng(X) (where Eng(X) means the English translation of X)
   Der Tisch gefallen Paul (source UR) → Paul like the table. (target UR)
3. target language grammar translates the target language UR into an actual target language sentence.
   Paul like the table → Paul likes the table

Things to note about transfer systems

- The transfer mechanism is essentially reversible; e.g., the gefallen rule works in both directions (at least in theory)
- Because we have a separate target language grammar, we are able to ensure that the rules of English apply: like → likes.
- Word order is handled differently than with transformers; the URs are essentially unordered.
- The underlying representation can be of various levels of abstraction — words, syntactic trees, meaning representations, etc.; we will talk about this with the translation triangle.

Caveat about reversibility

- It seems like reversible rules are highly desirable—and in general they are—but we may not always want reversible rules.
  - e.g., Dutch aanvangen should be translated into English as begin, but English begin should be translated into Dutch as beginnen.

Levels of abstraction

- There are differing levels of abstraction at which transfer can take place. So far we have looked at URs that represent only word information.
- We can do a full syntactic analysis, which helps us to know how the words in a sentence relate.
- Or we can do only a partial syntactic analysis, such as representing the dependencies between words.

Czech-English example

(8) Kaufman & Broad odmítla institucionální investory jmenovat.
   Kaufman & Broad declined institutional investors to name/identify ‘Kaufman & Broad refused to name the institutional investors.’

Example taken from Čmejrek, Čufín, and Havelka (2003).
- They find the base forms of words (e.g., obmítnout ‘to decline’ instead of odmítnout ‘declined’)
- They find which words depend on which other words and represent this in a tree (e.g., the noun investor depends on the verb odmítnout)
- This dependency tree is then converted to English (comparative grammar) and re-ordered as appropriate.
We are told what each word is translated as: text alignment
▶ We are not told what each word is translated as: use a bag of words

Interlingua example

<table>
<thead>
<tr>
<th>WOUND</th>
<th>MEANS</th>
</tr>
</thead>
<tbody>
<tr>
<td>gun</td>
<td>TENSE</td>
</tr>
<tr>
<td>kill</td>
<td>past</td>
</tr>
<tr>
<td>may</td>
<td>may</td>
</tr>
<tr>
<td>person</td>
<td>first</td>
</tr>
<tr>
<td>speaker</td>
<td>number</td>
</tr>
<tr>
<td>sg</td>
<td>gender</td>
</tr>
<tr>
<td>ACTION</td>
<td></td>
</tr>
<tr>
<td>sheriff</td>
<td>yes</td>
</tr>
<tr>
<td>PERSON</td>
<td>third</td>
</tr>
<tr>
<td>number</td>
<td>singular</td>
</tr>
<tr>
<td>HUMAN</td>
<td>yes</td>
</tr>
<tr>
<td>animate</td>
<td>yes</td>
</tr>
<tr>
<td>NOUN-TYPE</td>
<td>kind of job</td>
</tr>
<tr>
<td>is a kind of officer</td>
<td></td>
</tr>
</tbody>
</table>

Interlingual problems
▶ What exactly should be represented in the interlingua?
  ▶ e.g., English corner → Spanish rincon = ‘inside corner’ or esquina = ‘outside corner’
▶ A fine-grained interlingua can require extra (unnecessary) work:
  ▶ e.g., Japanese distinguishes older brother from younger brother, so we have to disambiguate English brother to put it into the interlingua. Then, if we translate into French, we have to ignore the disambiguation and simply translate it as frère, which simply means ‘brother’.

Using frequency (statistical methods)
▶ We can look at how often a source language word is translated as a target language word, i.e., the frequency of a given translation, and choose the most frequent translation.
▶ But how can we tell what a word is being translated as?
  There are two different cases:
  ▶ We are told what each word is translated as: text alignment
  ▶ We are not told what each word is translated as: use a bag of words

Sentence alignment
▶ sentence alignment = determine which source language sentences align with which target language ones (what we assumed in the bag of words example).
▶ Intuitively easy, but can be difficult in practice since different languages have different punctuation conventions.
**Different word alignments**

- One word can map to one word or to multiple words. Likewise, sometimes it is best for multiple words to align with multiple words.
- English-Hungarian examples:
  - one-to-one: well = jól
  - one-to-many: round = kör alakú
  - many-to-one: to play the guitar = gitárzók
  - many-to-many: even though = még ha ... is (‘even if ... also’)

**Calculating probabilities**

- With word alignments, it is relatively easy to calculate probabilities.
- E.g., What is the probability that run translates as rennen in German?
  1. Count up how many times run appears in the English part of your bi-text. E.g., 500 times
  2. Out of all those times, count up how many times it was translated as (i.e., aligns with) rennen. E.g., 275 (out of 500) times
  3. Divide to get a probability: 275/500 = 0.55, or 55%

**Word alignment difficulties**

- Knowing how words align in the training data will not tell us how to handle the new data we see.
  - We may have many cases where fool is aligned with the Spanish engañar = ‘to fool’
  - But we may then encounter a fool, where the translation should be tonto (male) or tonta (female)
- So, word alignment only helps us get some frequency numbers; we still have to do something intelligent with them.

**Word alignment difficulties (cont.)**

- Sometimes it is not even clear that word alignment is possible.
  - (9) Kati fotóz.
  - ‘Kati is a photographer.’
- What does an alignment mean?
  - In cases like this, a word can be mapped to a “null” element in the other language.

**The “bag of words” method**

- What if we’re not given word alignments?
- How can we tell which English words are translated as which German words if we are only given an English text and a corresponding German text?
  - We can treat each sentence as a bag of words – unordered collection of words.
    - If word A appears in a sentence, then we will record all of the words in the corresponding sentence in the other language as appearing with it.

**Example for bag of words method**

**Calculating probabilities: sentence 1**

So, for He in He speaks Hungarian well/O jól beszélnyi magyarul, we do the following:

1. Count up the number of Hungarian words: 4
2. Assign each word equal probability of translation: $1/4 = 0.25$, or 25%.

**Calculating probabilities: sentence 2**

If we also have He is a photographer/O fotós., then for He, we do the following:

1. Count up the number of possible translation words: 4 from the first sentence, 2 from the second = 6 total.
2. Count up the number of times O is the translation = 2 times out of 6 = $1/3 = 0.33$, or 33%.

Every other word has the probability $1/6 = 0.17$, or 17%, so On is clearly the best translation for O.

**Example for bag of words method (cont.)**

**Calculating probabilities: sentence 2**

If we also have He is a photographer/O fotós., then for He, we do the following:

1. Count up the number of possible translation words: 4 from the first sentence, 2 from the second = 6 total.
2. Count up the number of times O is the translation = 2 times out of 6 = $1/3 = 0.33$, or 33%.

Every other word has the probability $1/6 = 0.17$, or 17%, so On is clearly the best translation for O.

**What makes MT hard?**

- We’ve seen how MT systems can work, but MT is a very difficult task because languages are vastly different. They differ:
  - Lexically: In the words they use
  - Syntactically: In the constructions they allow
  - Semantically: In the way meanings work
  - Pragmatically: In what readers take from a sentence.

In addition, there is a good deal of real-world knowledge that goes into a translation.
Words can be **lexically ambiguous** = have multiple meanings.

- bank can be a financial institution or a place along a river.
- can be a cylindrical object, as well as the act of putting something into that cylinder (e.g., John cans tuna), as well as being a word like must, might, or should.

⇒ We have to know which meaning before we translate.

### How words divide up the world (lexical issues)

Words don’t line up exactly between languages.

- sofa and couch are synonyms (mean the same thing)
- sofa is a hyponym (more specific term) of furniture
- furniture is a hypernym (more general term) of sofa

### Synonyms

Often we find **synonyms** between two languages (as much as there are synonyms within a language):

- English book = Hungarian könyv
- English music = German Musik

But words don’t always line up exactly between languages.

### Venn diagram of semantic overlap

<table>
<thead>
<tr>
<th>paw</th>
<th>etape</th>
<th>journey</th>
</tr>
</thead>
<tbody>
<tr>
<td>patte</td>
<td>jambe</td>
<td>pied</td>
</tr>
<tr>
<td>animal</td>
<td>chair</td>
<td>human</td>
</tr>
</tbody>
</table>

### Hyponyms and Hypernyms

- English hypernyms = words that are more general in English than in their counterparts in other languages
  - English know is rendered by the French savoir (‘to know a fact’) and connaître (‘to know a thing’)
  - English library is German Bücherei if it is open to the public, but Bibliothek if it is intended for scholarly work.
- English hypernyms = words that are more specific in English than in their foreign language counterparts.
  - The German word Berg can mean either hill or mountain in English.
  - The Hungarian word láb can mean either leg or foot.

### Light verbs

Some verbs carry little meaning, so-called **light verbs**

- French *faire une promenade* is literally ‘make a walk,’ but it has the meaning of the English *take a walk*
- Dutch *e’en poging doen* ‘do an attempt’ means the same as the English *make an attempt*

### Idioms

And we often face **idioms** = expressions whose meaning is not made up of the meanings of the individual words.

- e.g., English kick the bucket
  - approximately equivalent to the German ins Gras beißen (‘bite into the grass’)
  - but we might want to translate it as sterben (*die*)
  - and we want to treat it differently than kick the table...
Taboo words

There are **taboo words** = words which are "forbidden" in some way or in some circumstances (i.e., swear/curse words)

- You, of course, know several English examples. Note that the literal meanings of these words lack the emotive impact of the actual words.
- Other languages/cultures have different taboos: often revolving around death, body parts, bodily functions, disease, and religion.
- e.g., The word 'skin' is taboo in a Western Australian (Aboriginal) language (http://www.aija.org.au/online/ICABenchbook/BenchbookChapter5.pdf)

Imagine encountering the word 'skin' in English and translating it without knowing this.

Structure and word order differences

- Word order (and syntactic structure) differs across languages.
- E.g., in English, we have what is called a subject-verb-object (SVO) order, as in (10).

  (10) John punched Bill, **SUBJECT VERB OBJECT**

- In contrast, Japanese is SOV. Arabic is VSO. Dyirbal (Australian aboriginal language) has free(r) word order.
- MT systems have to account for these differences.

More on word order differences

- Sometimes things are conceptualized differently in different languages, e.g.:

  a. My name is Adriane.
  b. Ich heiße Adriane. (German)
  c. Je m’appelle Adriane. (French)
  d. I myself call Adriane.
  e. Engem Adriennek hívnak. (Hungarian)
  f. Me Adriane they call

- Words don’t really align here.

How syntactic grouping and meaning relate (Syntax/Semantics)

Even within a language, there are syntactic complications. We can have **structural ambiguities** = sentences where there are multiple ways of interpreting it:

(11) a. My name is Adriane.
    b. Ich heiße Adriane. (German)
    c. Je m’appelle Adriane. (French)
    d. I myself call Adriane.
    e. Engem Adriennek hívnak. (Hungarian)
    f. Me Adriane they call

- Do we attempt to translate only one interpretation? Or do we try to preserve the ambiguity in the target language?

How language is used (Pragmatics)

Translation becomes even more difficult when we try to translate something in context.

- Thank you is usually translated as merci in French, but it is translated as s’il vous plaît ‘please’ when responding to an offer.
- Can you drive a stick-shift? could be a request for you to drive my manual transmission automobile, or it could simply be a request for information about your driving abilities.

Real-world knowledge

- Sometimes we have to use **real-world knowledge** to figure out what a sentence means.

  (13) Put the paper in the printer. Then switch it on.

  - We know what it refers to only because we know that printers, not paper, can be switched on.

Ambiguity resolution

- If the source language involves ambiguous words/phrases, but the target language does not have the same ambiguity, we have to resolve ambiguity before translation.
- e.g., the hyponyms/hypernyms we saw before.
- But sometimes we might want to preserve the ambiguity, or note that there was ambiguity or that there are a whole range of meanings available.
- In the Bible, the Greek word hyper is used in 1 Corinthians 15:29; it can mean ‘over’, ‘for’, ‘on behalf of’, and so on. How you treat it affects how you treat the theological issue of salvation of the dead. So, people care deeply about how you translate this word, yet it is not entirely clear what English meaning it has.

Evaluating MT systems

- We’ve seen some translation systems and we know that translation is hard.
- The question now is: How do we evaluate MT systems, in particular for use in large corporations as likely users?
  - How much change in the current setup will the MT system force?
  - Translator tasks will change from translation to updating the MT dictionaries and post-editing the results.
  - How will it fit in with word processors and other software?
  - Will the company selling the MT system be around in the next few years for support and updates?
  - How fast is the MT system?
  - How good is the MT system (quality)?
Intelligibility

Intelligibility Scale (from Arnold et al., 1994)

1. The sentence is perfectly clear and intelligible. It is grammatical and reads like ordinary text.
2. The sentence is generally clear and intelligible. Despite some inaccuracies or infelicities of the sentence, one can understand (almost) immediately what it means.
3. The general idea of the sentence is intelligible only after considerable study. The sentence contains grammatical errors and/or poor word choices.
4. The sentence is unintelligible. Studying the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable.

Further reading

Some of the examples are adapted from the following books:


Evaluating quality

- **Intelligibility** = how understandable the output is
- **Accuracy** = how faithful the output is to the input
- **Error analysis** = how many errors we have to sort through (and how do the errors affect intelligibility & accuracy)
- **Test suite** = a set of sentences that our system should be able to handle