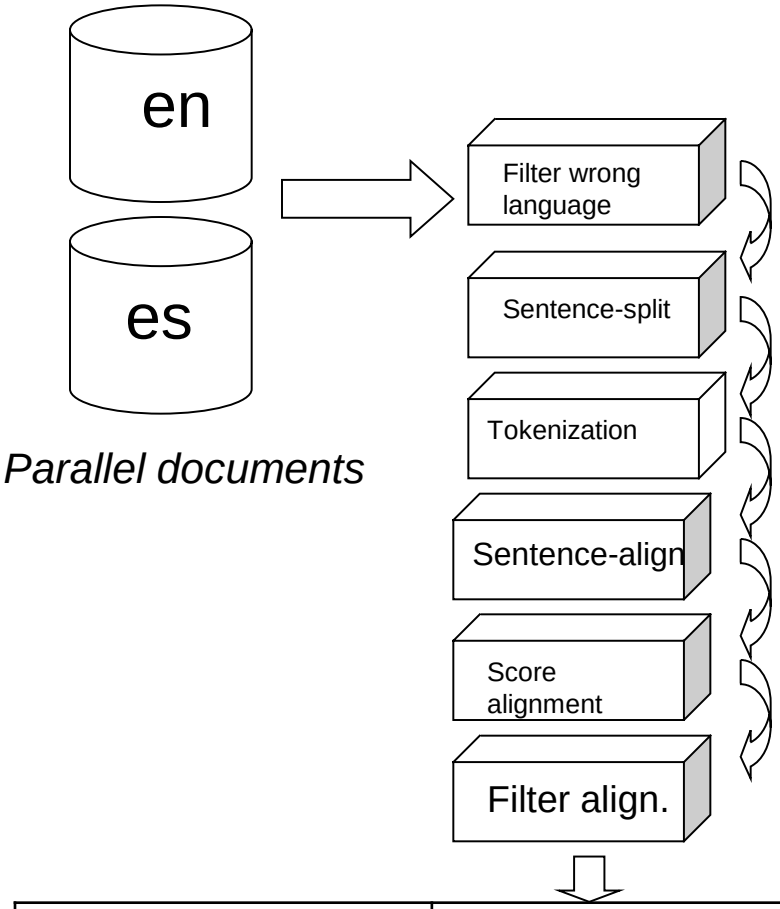
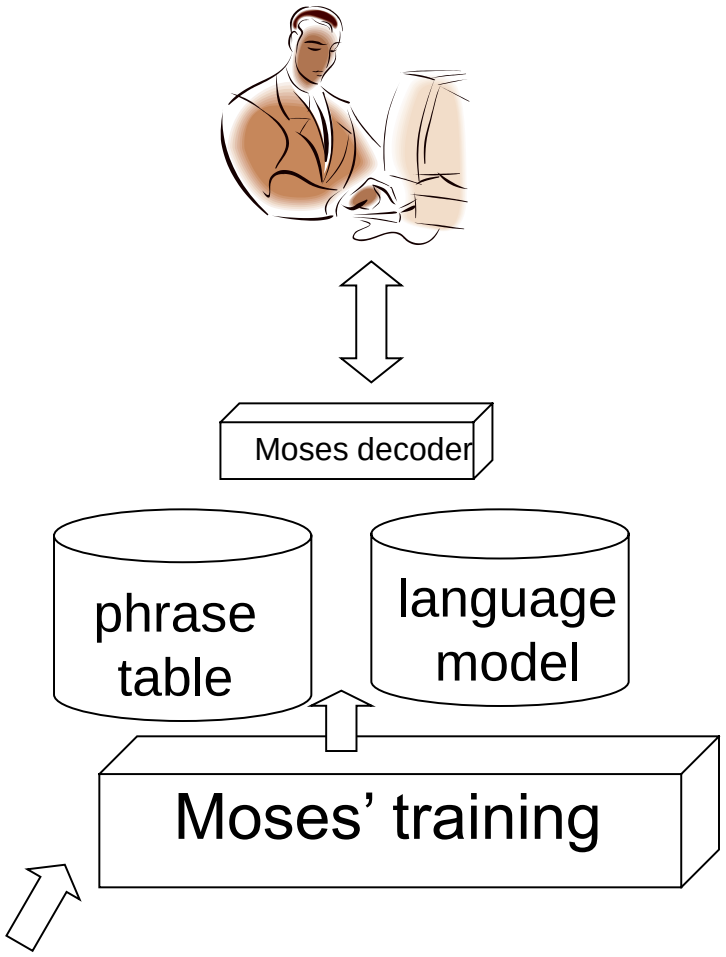
Lecture 5.4: MT preparation steps

How to prepare your texts for MT,
pre and post processing

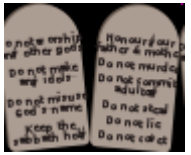
SMT lifecycle



En	Es
Strengthening of forum for human dignity : legal aid	Fortalecimiento del foro para la dignidad humana – asistencia jurídica
must respect all aspects of human dignity	debe respetar todos los aspectos de la dignidad humana
should fully respect human	se deben respetar plenamente la



Moses input format



- 2 text files (utf-8), called bitext
- Each line is a tokenized sentence (hopefully short <80w)

  Each line in first file has to be the translation of corresponding line in second file
- No id, no classification etc.

Tokenization

- Moses translates tokens (usually words separated by space)

  Tony ≠ Tony's ≠ Tony. ≠ tony ≠ Tony, ≠ Tony's

  Japanese/Chinese tokenization needed

Tokenizer, options

For example, our tokenizer in WIPO does:

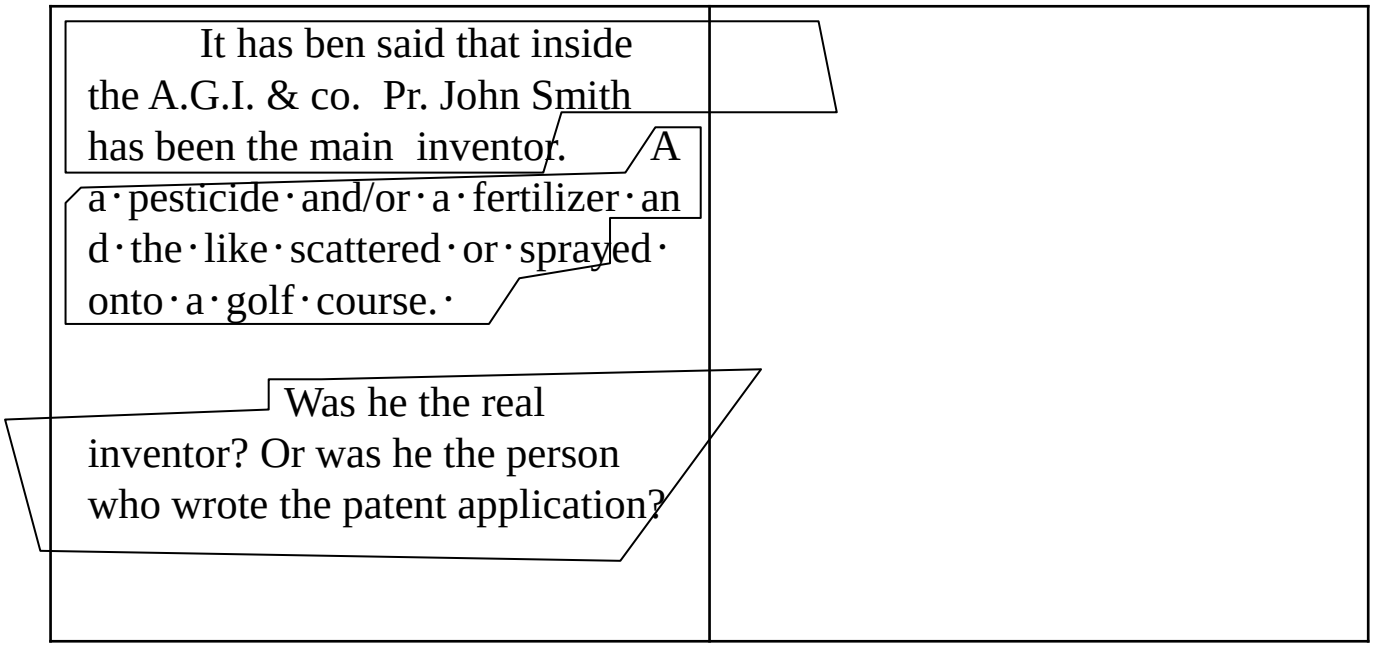
- Extract email address as one single token
- Extract full URLs as one single token (tagged as HOST)
- Recognizes Xml/Html tags as one token
- Reduces acronyms without dots (“U.N.” become “UN”)
- French apostrophe: “jusqu’alors” becomes “jusqu’_alors”
- English possessive case: “UN’s opinion” becomes “UN_’s_opinion”, (but “al-Ma'sara” becomes “al_-_ma’sara”)
- Special recasing: “The New York UN HeadQuarter” ➔ “the New York UN headquarter”
- Groups references to figures as one token (eg. “(1)” not “(_1_)”)
- Recognize "and / or" as one single token
- Special case of greek letters "α- and/or β-amino acid" becomes "alpha - and/or beta - amino acid “
- Harmonize language specific diacritics? Latin “ae” => ae, delete Arabic short vowels?

Tokenizer(2)

- Chinese tokenizer: smartcn (small adaptations)
  - “演奏音乐的娱乐装置” => “演奏 / 音乐 / 的 / 娱乐 / 装置”
- German decompounding
  - “Lebenshaltungskostenausgleich” ➔ “Lebens- haltungs- kosten- ausgleich”
- Arabic prefix splitting
  - splits prefixes "ال“,"وال“,"لل“, "بال“, "كال“
- Indic languages tokenization
  - http://anoopkunchukuttan.github.io/indic_nlp_library/ python script
  - Be careful: the Moses’ tokenizer does not handle properly Indic languages
  - For the labs, I quickly adapted the Moses tokenizer (Perl script), see Icon2016/labs/tokenizer.pl

Sentence splitting

- MT works with sentences, a document must be splited in sentences



Sentence splitting

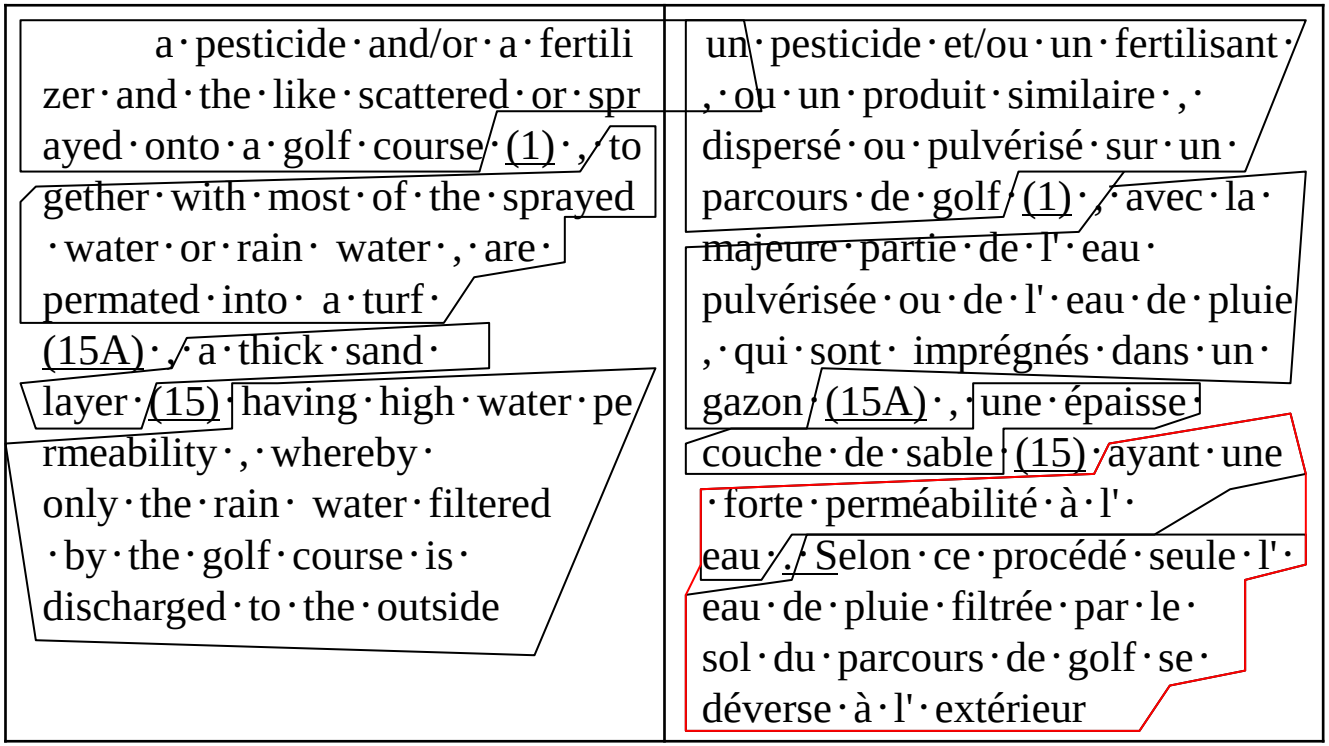
- Usually using full stop as a first approximation
  - However, abbreviations may also contains full stops.
  - E.g.: “As Pr. Smith, no. 123, 13 dec. 2015”
- Depends on the language
  - Chinese “。”, Devanagari “।”, Armenian “։”, Amharic “።”…
  - In Thai, a space is the equivalent of a full stop

Sentence alignment

- Documents may contain different sentence structure (sentences are not mapped one-to-one)
- Sentence alignment tools are needed
  - Count n. of words, find best 1-0,0-1,1-1,1-2,2-1 combinations to align [Brown et al.1991])
  - Use bilingual dictionary
  - Use cognates, names, numbers
  - See [Singh & Hussain 2005]

- More on http://www.statmt.org/survey/Topic/SentenceAlignment

Sentence alignment example



Document alignment

- Challenges with getting parallel sentences from comparable documents

[0009] Zur Lösung dieses Problems sieht die Erfindung vor, dass bei dem eingangs genannten Tonträger wenigstens eine der Justierdurchbrechungen auf der Oberseite des Tonträgers von einem nach oben aus der Oberseite vorstehenden Wulstring, der z. B. eingeprägt oder

To meet this object, the invention provides that in the sound carrie one of the adjustment perforations on the surface of the sound car annular bead which protrudes in upward direction from the surface deep-drawn, for example. Said annular bead may either be located perforation or surround it a small distance of about 1.0 to 1.5 mm, forming a flat shoulder which contributes to defining the perforatio have a height of 0.2 to 0.3 mm, especially of 0.2 mm, and a width o 1.2 mm.

It is advisable to provide the two perforations in the lower edge such an annular bead. Such an annular bead does not only make th they also render the alignment of the feet of the sound reproducin practically fall into the larger annular bead and slip into the precise automatically. Such a perforation surrounded by an annular bead c funnel.

Of course, all the perforations in a sound carrier can be provide prevent an unintentional dislocation of the sound reproducing devi interruptions of a reproduction while studying educational informa sound reproducing device has to be actuated frequently also during for interruptions or for a repeated reproduction which naturally inv dislocation.

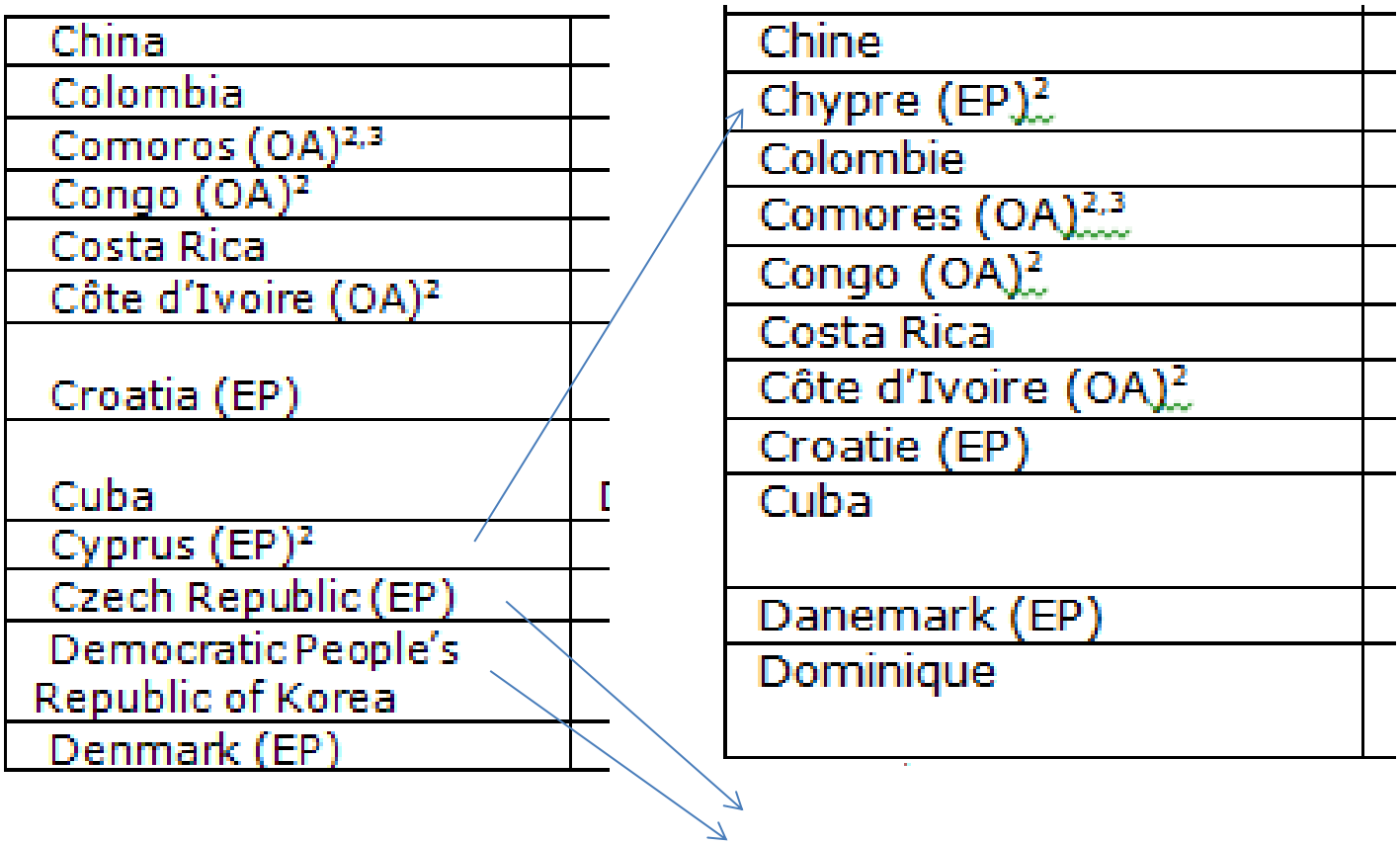
As an alternative to this solution, the two lower adjustment per and elongated in the direction away from the other adjustment per or in the direction of the same toward the inside and surround the inserted into them with large clearance by means of the wide regio that said aligning foot is movable to abut the narrow region and th is/are in a position to engage the other adjustment perforation/per plane guiding funnel is formed in this manner which passes the sou placed in position into the round perforations of the sound carrier.

The walls of the enlarged adjustment perforations should have the side facing toward or away from the other adjustment perforat inserted. As a result, when centered, the feet are in each case in th which they would also be if all the perforations were, as is known, perforations not made larger should have the common diameter, The enlarged perforations have by nearly their halves the same co widen and are extended to reach a larger transverse dimension of adjustment perforations are open toward the underside of the sou adjustment perforations of which one is circular and the other fun the alignment. Nevertheless, mostly four perforations are provide previously, of which two are widened to funnel shape.

surround it a small distance of about 1.0 to 1.5 mm , especially 1.2 mm , thus forming a flat shoulder which contributesto defining the perforation	1-1	99.28%	einen geringen Abstand von etwa 1,0 bis 1,5 mm , insbesondere 1,2 mm , und bildet so einen flachen Absatz , was zur Präzisierung der Durchbrechung beiträgt
the annular bead should have a height of 0.2 to 0.3 mm , especially of 0.2 mm , and a width ofabout 1 -1.5 mm , especially 1.2 mm . it is advisable to provide the two perforations in the lower edge of the sound carrier with such an annular bead	1-2	48.85%	Der Wulstring sollte eine Höhe von 0,2 bis 0,3 mm , insbesondere von 0,2 mm , und eine Breite von ca . 1 - 1,5 mm , insbesondere 1,2 mm haben . Zweckmässigerweise werden zwei Durchbrüche am unteren Rand des Tonträgers mit diesem Wulstring ausgestattet
such anannular bead does not only make the perforations better visible , they also render the alignment of the feet of the sound reproducingdevice easier because they practically fall into the larger annular bead and slip into the precise adjustment perforationsautomatically	1-1	90.33%	Durch diesen Wulstring sind die Durchbrüche nicht nur besser sichtbar , sie erleichtern auch das Justieren der Füsse des Tonwiedergabegerätes , da die Justierfüsse in dem grösseren Wulstring praktisch hineinfallen und zwangsweise in die präzisen Justierdurchbrechungen schlüpfen
such a perforation surrounded by an annular bead can actually be compared to a funnel	1-1	21.93%	Man kann diese von einem Wulstring umgebenen Durchbrüche praktisch mit einem Trichter vergleichen .
of course , all the perforations in a sound carrier can be provided with beads	1-1	30.73%	Selbstverständlich können auch alle Durchbrechungen eines Tonträgers mit Wulsten versehen werden
they also prevent an unintentional dislocationof the sound reproducing device during possibly desired interruptions of a reproduction while studying educational information.This is important as the sound reproducing device has to be actuated frequently also during reproduction , for example forinterruptions or for a repeated reproduction which naturally involves the danger of dislocation	1-2	50.66%	Diese Wülste verhindern auch das unbeabsichtigte Verrücken des Tonwiedergabegerätes bei evtl . gewollten Unterbrechungen der Wiedergabe beim Studium von Lerntexten Das ist wichtig , weil das Tonwiedergabegerät auch während der Wiedergabe häufig , beispielsweise für Unterbrechungen oder Wiedergabewiederholungen zu betätigen ist und dabei an sich die Gefahr des Verrückens besteht .
as an alternative to this solution , the two lower adjustment perforations may be widened and elongated in the direction awayfrom the other adjustment perforations toward the outside or in the direction of the same toward the inside and surround therespective aligning foot to be inserted into them with large clearance by means of the wide region created in this manner , so that said aligning foot is movable to abut the narrow region and the other aligning foot / feet is / are in a position to engagethe other adjustment perforation / perforations in this position	1-1	100%	Alternativ zu dieser Lösung kann auch ein Teil der Justierdurchbrechungen , z . B. zwei untere Justierdurchbrechungen , in Richtung von den anderen Justierdurchbrechungen nach aussen weg oder zu diesen nach innen hin verbreitert und verlängert ausgebildet sein und den in sie einzuführenden Justierfuss mit dem so geschaffenen breiten Bereich mit grossem Spiel so umschliessen , dass dieser Justierfuss zur Anlage an den engen Bereich verschieblich ist und die anderen Justierfüsse in dieser Lage in die anderen Justierdurchbrechungen , die die Justierfüsse im engsten Bereich mit geringem Spiel umschliessen , einzugreifen vermögen
a plane guiding funnel is formed in this manner which passesthe sound reproducing device placed in position into the round perforations of the sound carrier	1-1	99.86%	So ist ein ebener Führungstrichter gebildet , der das aufgesetzte Tonwiedergabegerät in die runden Durchbrechungen des Tonträgers leitet
	0-1		Zusammen mit den dann zum Eingriff kommenden anderen Justierfüssen ist das Tonwiedergabegerät dann lagestabil gehalten .
	0-1		Ein älterer bekannter Vorschlag (US Patentschrift 4,298,967) ist von der Praxis als ungeeignet verworfen worden , der im Mittelpunkt der Tonrille eines folienartigen , dünnen Tonträgers eine schuhartige zur Mitte und nach unten sich verengende Vertiefung vorsah , in die ein zentraler Justierstift des Tonwiedergabegerätes bis zum Grund in die Justierstellung geführt wird , in diese aber wegen Fehlen der Arretierungsmöglichkeit nicht lagestabil festgehalten ist
	0-1		Ausserdem verdeckt das Tonwiedergabegerät jegliche Sicht zur schuhartigen Vertiefung .
the walls of the enlarged adjustment perforations should have the same configuration on the side facing toward or away fromthe other adjustment perforations as the aligning feet to be inserted	1-1	57.12%	Die Wände der vergrösserten Justierdurchbrechungen sollten auf der den anderen Justierdurchbrechungen zugewandten oder abgewandten Seite die gleiche Gestalt haben wie die einzusetzenden Justierfüsse
as a result , when centered , the feet are in each casein the exactly defined			Dadurch befinden sich die Füsse lagestabil in der zentrierten Lage jeweils an den genau

Text specificities

- Challenges with parallel sentences from documents (here lists sorted alphabetically)



China
Colombia
Comoros (OA)[2,3]
Congo (OA)[2]
Costa Rica
Côte d'Ivoire (OA)[2]
Croatia (EP)
Cuba
Cyprus (EP)[2]
Czech Republic (EP)
Democratic People's Republic of Korea
Denmark (EP)

Chine
Chypre (EP)[2]
Colombie
Comores (OA)[2,3]
Congo (OA)[2]
Costa Rica
Côte d'Ivoire (OA)[2]
Croatie (EP)
Cuba
Danemark (EP)
Dominique

Clean long lines

- Filter out sentences having more than 80 words
- Ratio (# src words/ # dst words) more than 9

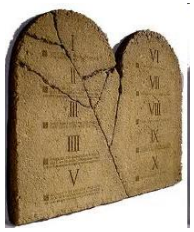
```
perl ~/mosedecoder/script/bin/clean-corpus-n.perl training en es
   cleaned 1 80
```

Word alignment (giza) complexity is linear according to the number of sentences, however it is quadratic according to the number of tokens per sentence

Trainset / devset / testset

- Trainset: Set of sentences used by your machine to learn (& language model generation)
- Devset: Set of sentences used to optimize the model parameters (Moses “MERT”)
- Testset: Set of sentences used for comparison between automatic translation and human translation (e.g. BLEU)

Trainset / devset / testset few commandments



- Never mix trainset and testset (Even for the language model)
- Always test on held-out documents
- Try to have a representative test/dev set (In some context: translate new documents)
  - Test on newest documents
  - Develop on newest documents
- Use appropriate evaluation measure (human evaluation, BLEU etc.), see lecture 6