European Language Resource Coordination
Connecting Europe Facility

“What Data Is Needed? Why ?”

Andrea Lösch, DFKI
Andrejs Vasiļjevs, Tilde

Credits: Khalid Choukri, ELRA

European Language Resource Coordination
Connecting Europe Facility

What is needed for MT?

- From the previous sessions, we have seen the predominant approach of data-driven paradigm
  - MT systems learn from existing data
  - Focus for ELRC: Data in all languages (EU/CEF)

- How are Language Resources produced?
  - from documents and data to Language Resources

European Language Resource Coordination
Connecting Europe Facility

What counts as data for MT?

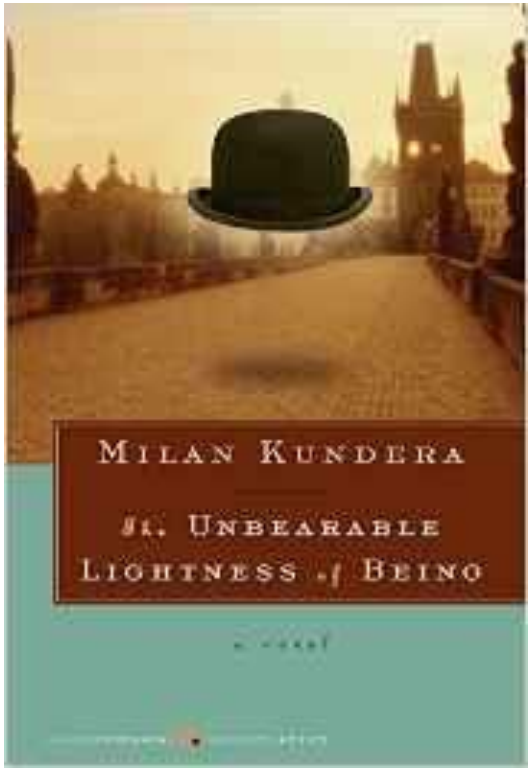
- Anything that contains “words”, preferences for “sentences”, even for sentences expressed in multiple languages, e.g.
  - Reports,
  - Speeches,
  - Contents on web pages,
  - Brochures, etc.

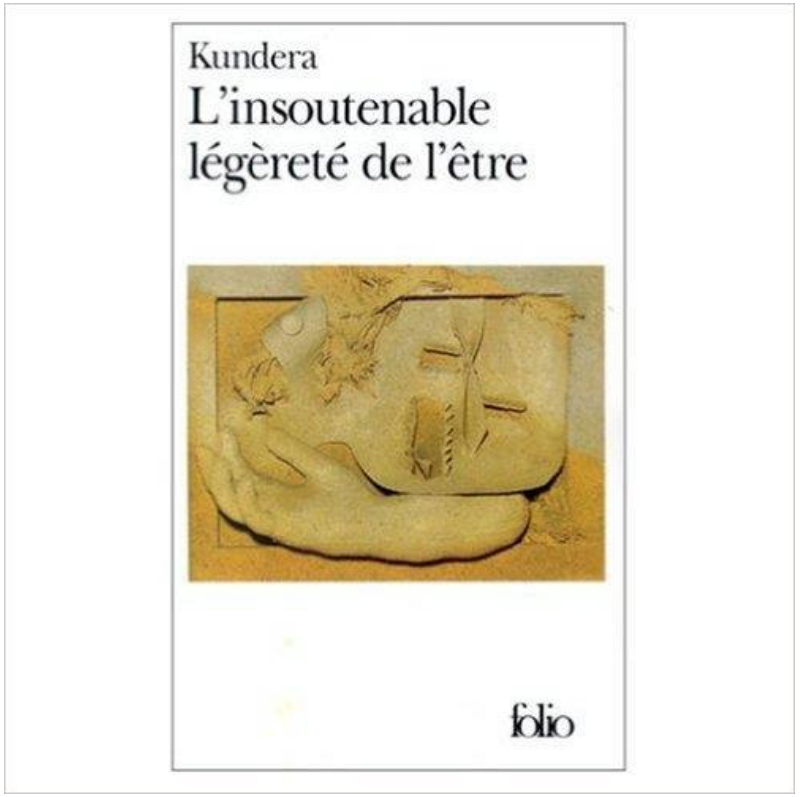
- Bags of “words”, “sentences”, multiple bags

What counts as data for MT?

European Language Resource Coordination
Connecting Europe Facility

What types of data? Translations





European Language Resource Coordination
Connecting Europe Facility

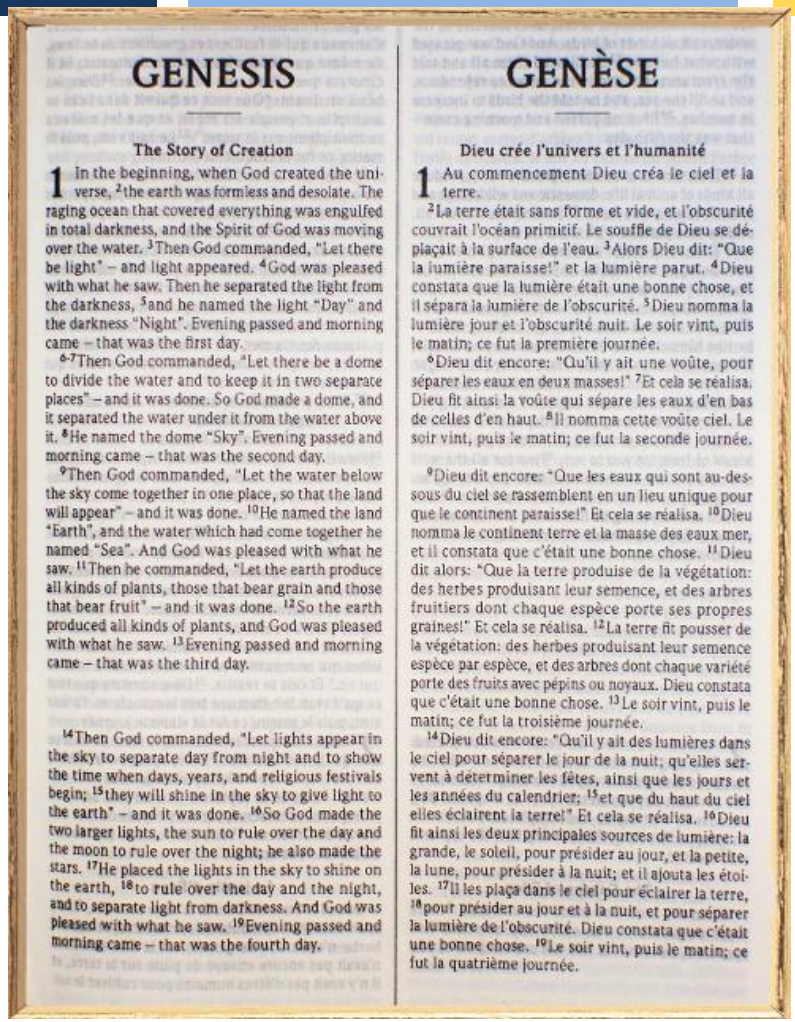
What types of data? "Aligned" Translations

English

French

European Language Resource Coordination
Connecting Europe Facility

What types of data?
"Aligned" Translation



GENESIS

The Story of Creation

1 In the beginning, when God created the universe, 2the earth was formless and desolate. The raging ocean that covered everything was engulfed in total darkness, and the Spirit of God was moving over the water. 3Then God commanded, "Let there be light" – and light appeared. 4God was pleased with what he saw. Then he separated the light from the darkness, 5and he named the light "Day" and the darkness "Night". Evening passed and morning came – that was the first day.

6-7Then God commanded, "Let there be a dome to divide the water and to keep it in two separate places" – and it was done. So God made a dome, and it separated the water under it from the water above it. 8He named the dome "Sky". Evening passed and morning came – that was the second day.

9Then God commanded, "Let the water below the sky come together in one place, so that the land will appear" – and it was done. 10He named the land "Earth", and the water which had come together he named "Sea". And God was pleased with what he saw. 11Then he commanded, "Let the earth produce all kinds of plants, those that bear grain and those that bear fruit" – and it was done. 12So the earth produced all kinds of plants, and God was pleased with what he saw. 13Evening passed and morning came – that was the third day.

14Then God commanded, "Let lights appear in the sky to separate day from night and to show the time when days, years, and religious festivals begin; 15they will shine in the sky to give light to the earth" – and it was done. 16So God made the two larger lights, the sun to rule over the day and the moon to rule over the night; he also made the stars. 17He placed the lights in the sky to shine on the earth, 18to rule over the day and the night, and to separate light from darkness. And God was pleased with what he saw. 19Evening passed and morning came – that was the fourth day.

GENÈSE

Dieu crée l'univers et l'humanité

1 Au commencement Dieu créa le ciel et la terre.

2La terre était sans forme et vide, et l'obscurité couvrait l'océan primitif. Le souffle de Dieu se déplaçait à la surface de l'eau. 3Alors Dieu dit: "Que la lumière paraisse!" et la lumière parut. 4Dieu constata que la lumière était une bonne chose, et il sépara la lumière de l'obscurité. 5Dieu nomma la lumière jour et l'obscurité nuit. Le soir vint, puis le matin; ce fut la première journée.

6Dieu dit encore: "Qu'il y ait une voûte, pour séparer les eaux en deux masses!" 7Et cela se réalisa. Dieu fit ainsi la voûte qui sépare les eaux d'en bas de celles d'en haut. 8Il nomma cette voûte ciel. Le soir vint, puis le matin; ce fut la seconde journée.

9Dieu dit encore: "Que les eaux qui sont au-dessous du ciel se rassemblent en un lieu unique pour que le continent paraisse!" Et cela se réalisa. 10Dieu nomma le continent terre et la masse des eaux mer, et il constata que c'était une bonne chose. 11Dieu dit alors: "Que la terre produise de la végétation: des herbes produisant leur semence, et des arbres fruitiers dont chaque espèce porte ses propres graines!" Et cela se réalisa. 12La terre fit pousser de la végétation: des herbes produisant leur semence espèce par espèce, et des arbres dont chaque variété porte des fruits avec pépins ou noyaux. Dieu constata que c'était une bonne chose. 13Le soir vint, puis le matin; ce fut la troisième journée.

14Dieu dit encore: "Qu'il y ait des lumières dans le ciel pour séparer le jour de la nuit; qu'elles servent à déterminer les fêtes, ainsi que les jours et les années du calendrier; 15et que du haut du ciel elles éclairent la terre!" Et cela se réalisa. 16Dieu fit ainsi les deux principales sources de lumière: la grande, le soleil, pour présider au jour, et la petite, la lune, pour présider à la nuit; et il ajouta les étoiles. 17Il les plaça dans le ciel pour éclairer la terre, 18pour présider au jour et à la nuit, et pour séparer la lumière de l'obscurité. Dieu constata que c'était une bonne chose. 19Le soir vint, puis le matin; ce fut la quatrième journée.

European Language Resource Coordination
Connecting Europe Facility

What types of data?
"Aligned" Translation

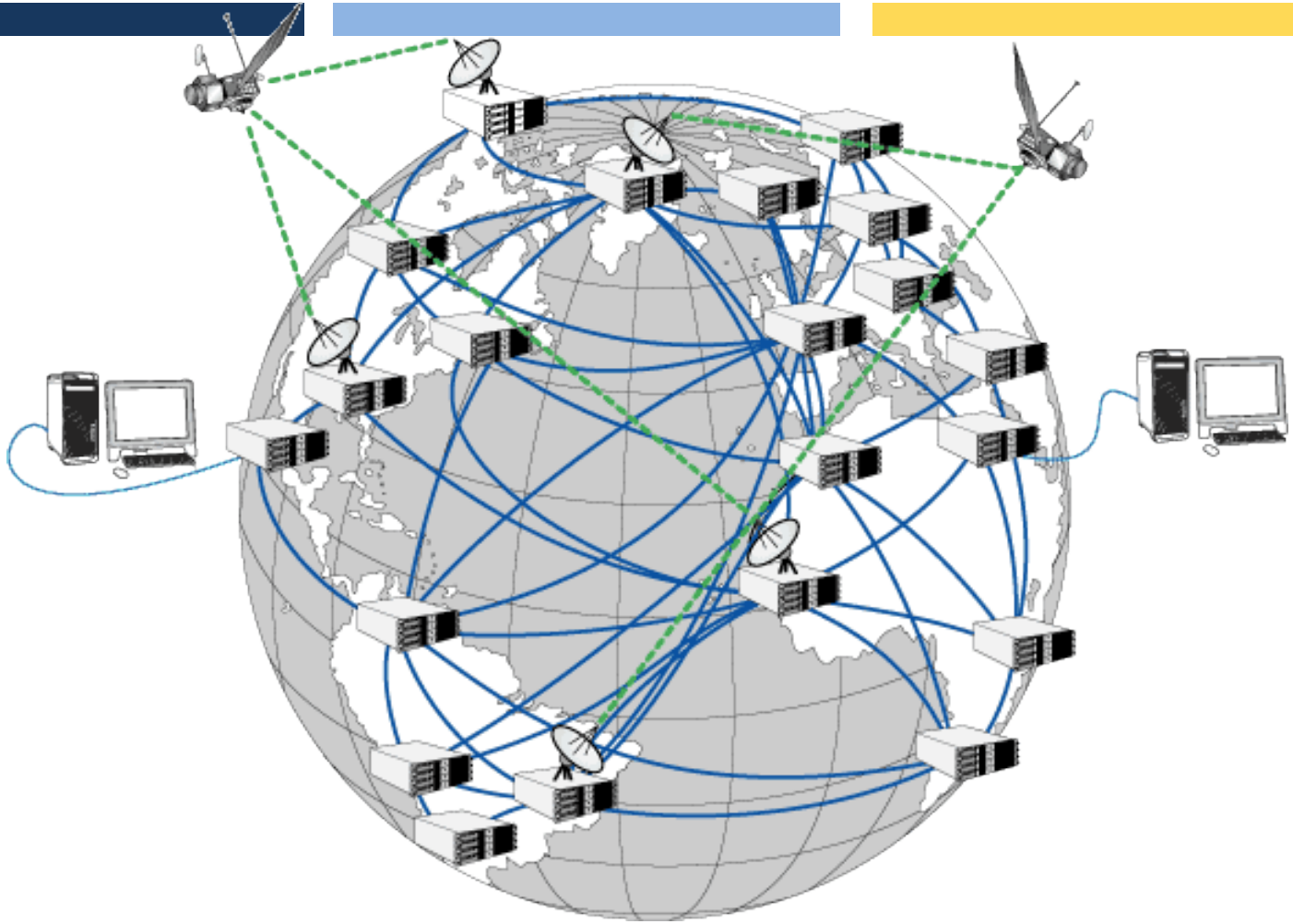
The Vikings were Scandinavian seafarers who lived in the ninth, tenth, and the beginning of the eleventh century, which is known as the Viking era. The Vikings were heathens and did not become Christian until around the year 1000. Their own gods were called the Æsir, and offerings were made to them at the blot, a kind of religious sacrificial holiday.	Die Wikinger waren skandinavische Seefahrer, die im 9., 10. und Anfang des 11. Jahrhunderts lebten, auch bekannt als Wikinger-Epoche. Die Wikinger waren Heiden und wurden erst um das Jahr 1000 zu Christen. Ihre eigenen Götter nannten sie Æsir, denen sie am Blot, einem religiösen Opfertag, Gaben darbrachten. Vier dieser Götter waren Tyr (oder Tiwaz), Odin (oder Wotan), Thor und Frigga, nach denen drei Wochentage benannt sind: Dienstag, Donnerstag und Freitag. Auch die Monate hatten ihre eigenen Namen, aber heutzutage benutzen die Skandinavier die römischen Namen für die Monate: Januar, Februar, März etc.
Four of these gods were Tyr (or Tiwaz), Odin (or Wotan), Thor, and Frigga, who have given their names to four of the days of the week: Tuesday, Wednesday, Thursday and Friday. The months had their own names as well, but now the Scandinavians use the Roman names for the months: January, February, March etc.	
Many Vikings sailed out into the world in their long-ships, or drekkar, as far as America and Constantinople. Their ships had relatively flat bottoms, so that they could sail near the coast and up shallow rivers. In the West they met Indians, and in the East they met Arabs. Out in the Atlantic they navigated by the stars, and in the year 1000 Leif Eriksson set foot on American soil, and forty years later, Ingvar the Wide-Traveled reached the southern shore of the Caspian sea. In this way, local kings had contact with lands which lay far away. In large areas of England Danish law held sway; that area was therefore called the Danelaw. In Constantinople, the emperor had a feared bodyguard composed of Vikings. Because of their distinctive axes, they were called "the Axe-bearing Barbarians."	Viele Wikinger segelten in ihren Langschiffen oder Drekkar hinaus in die Welt, bis nach Amerika und Konstantinopel. Ihre Schiffe hatten relativ flache Böden, so daß sie sich damit auch nahe der Küste und in seichten Flüssen bewegen konnten. Im Westen begegneten sie Indianern und im Osten Arabern. Auf dem Atlantik navigierten sie mit Hilfe der Sterne und im Jahr 1000 setzte Leif Eriksson seinen Fuß auf amerikanischen Boden, und vierzig Jahre später erreichte Ingvar, 'der Weitgereiste', die Südküste des Kaspischen Meeres. Auf diese Weise kamen einheimische Könige in Kontakt mit Ländern, die weit entfernt waren. In weiten Teilen Englands herrschte dänisches Gesetz. Diese Gebiete wurden deshalb Danelaw genannt. In Konstantinopel hielt sich der Herrscher eine gefürchtete Wikingergarde. Wegen ihrer typischen Streitäxte wurden sie die Axt-tragenden Barbaren genannt.
At home the Vikings lived relatively simply. They sowed rye in the fields and kept cows, which gave milk, pigs, for pork, and sheep, for wool. Those who lived along the coasts caught fish. They often lived in long-houses, which could house several families. Three or four brothers, for example, could live with their families together in one big house.	Zu Hause lebten die Wikinger recht einfach. Auf den Feldern kultivierten sie Roggen und sie hielten Kühe, die sie mit Milch versorgten, Schweine hielten sie wegen des Fleisches und Schafe für Wolle. Jene, die an der Küste lebten, fingen Fisch. Die Wikinger wohnten gewöhnlich in Langhäusern, die mehrere Familien beherbergen konnten. Drei oder vier Brüder konnten, zum Beispiel, zusammen mit ihren Familien in einem einzigen großen Haus leben.

Dictionaries / Terminologies / Ontologies

ID	FR	ES	EL
6905	abandon scolaire	abandono escolar	διακοπή της σχολικής φοίτησης
920	abats	despojo	παραπροϊόντα σφαγίων
1857	abattage d'animaux	sacrificio de animales	σφαγή ζώων
6621	abrogation	derogación	κατάργηση
5075	Abruzzes	Abruzos	Αβρουζία
5339	absentéisme	absentismo	συστηματική απουσία από την εργασία
5984	abstentionnisme	abstencionismo	αποχή
2	abus de confiance	abuso de confianza	απιστία
96	abus de droit	abuso de derecho	κατάχρηση δικαιώματος
	abus de pouvoir	abuso de poder	κατάχρηση εξουσίας
	accès à l'éducation	acceso a la educación	πρόσβαση στην εκπαίδευση
	accès à l'emploi	acceso al empleo	πρόσβαση στην αγορά εργασίας

European Language Resource Coordination
Connecting Europe Facility

What data format?
Internet & Digital Data



European Language Resource Coordination
Connecting Europe Facility

What format is needed? Digital textual data

European Language Resource Coordination
Connecting Europe Facility

How Language Resources (LR) are produced from data

- Let us see some examples of raw data (html with tables, pictures, etc.) and how they become LRs
  - Discover & identify sources
  - Clear IPR and Get the data (Download, harvest, crawl, …)
  - Clean the data (e.g. detect and remove the “boilerplate”, “templates”, pictures, html tags, etc., convert format)
  - Example of tools (Boilerpipe)
  - Document the data
  - Align the translations when identified and break into “sentences”
  - Compute some alignment confidence
  - Share