CL meets IfI

Johannes Graën

2014-07-08

Machine Translation

and

Large Text Corpora for Linguistic Research

Machine Translation

Machine Translation

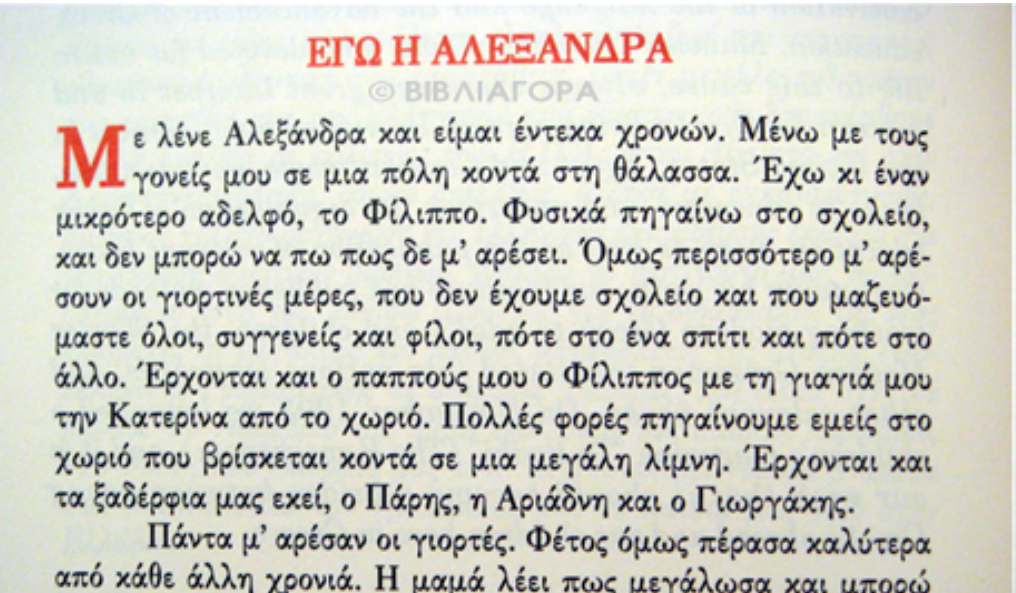
- Rule based vs. statistical machine translation (SMT)

Machine Translation

- Rule based vs. statistical machine translation (SMT)
- High quality translation vs. gist translation

Machine Translation

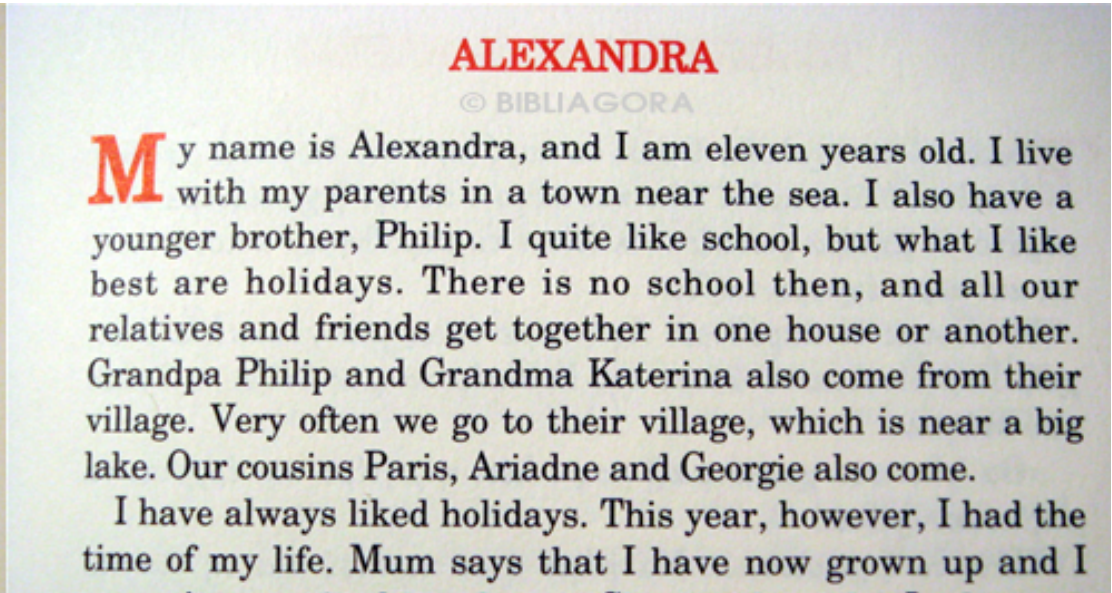
- Rule based vs. statistical machine translation (SMT)
- High quality translation vs. gist translation
- SMT based on parallel text corpora:



ΕΓΩ Η ΑΛΕΞΑΝΔΡΑ

Με λένε Αλεξάνδρα και είμαι έντεκα χρονών. Μένω με τους γονείς μου σε μια πόλη κοντά στη θάλασσα. Έχω κι έναν μικρότερο αδελφό, το Φίλιππο. Φυσικά πηγαίνω στο σχολείο, και δεν μπορώ να πω πως δε μ' αρέσει. Όμως περισσότερο μ' αρέσουν οι γιορτινές μέρες, που δεν έχουμε σχολείο και που μαζευόμαστε όλοι, συγγενείς και φίλοι, πότε στο ένα σπίτι και πότε στο άλλο. Έρχονται και ο παππούς μου ο Φίλιππος με τη γιαγιά μου την Κατερίνα από το χωριό. Πολλές φορές πηγαίνουμε εμείς στο χωριό που βρίσκεται κοντά σε μια μεγάλη λίμνη. Έρχονται και τα ξαδέρφια μας εκεί, ο Πάρης, η Αριάδνη και ο Γιωργάκης.

Πάντα μ' αρέσαν οι γιορτές. Φέτος όμως πέρασα καλύτερα από κάθε άλλη χρονιά. Η μαμά λέει πως μεγάλωσα και μπορώ



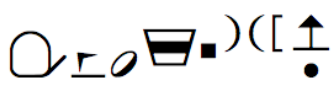
ALEXANDRA

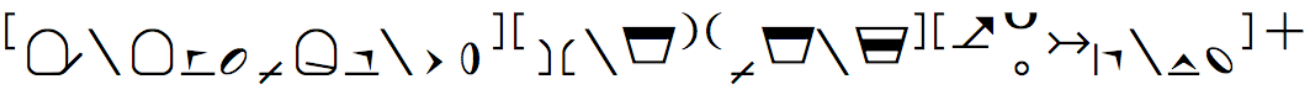
My name is Alexandra, and I am eleven years old. I live with my parents in a town near the sea. I also have a younger brother, Philip. I quite like school, but what I like best are holidays. There is no school then, and all our relatives and friends get together in one house or another. Grandpa Philip and Grandma Katerina also come from their village. Very often we go to their village, which is near a big lake. Our cousins Paris, Ariadne and Georgie also come.

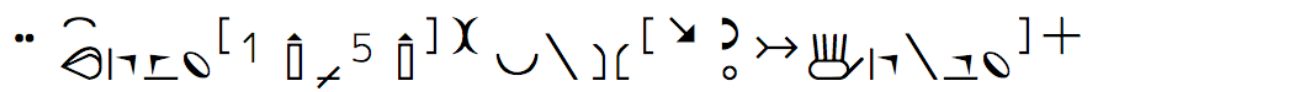
I have always liked holidays. This year, however, I had the time of my life. Mum says that I have now grown up and I

Machine Translation

Sign Language





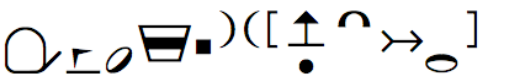


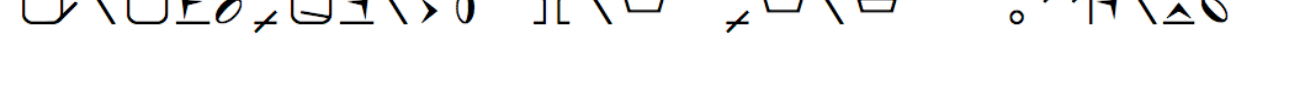
Machine Translation

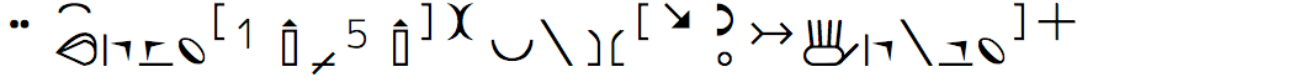
Sign Language

Machine Translation

Sign Language







Machine Translation

Sign Language

Machine Translation
Sign Language

Machine Translation
Sign Language

Machine Translation

Coreferenced German compounds

Die Originalauswertung wurde in den Zwischenmassstab 1:200 reduziert, worauf das Bundesamt (SMT: *office fédéral*) [...]

Machine Translation

Coreferenced German compounds

Die Originalauswertung wurde in den Zwischenmassstab 1:200 reduziert, worauf das Bundesamt (SMT: *office fédéral*) [...]

SRC Nur dieses Amt war in der Lage, [...]

Machine Translation

Coreferenced German compounds

Die Originalauswertung wurde in den Zwischenmassstab 1:200 reduziert, worauf das Bundesamt (SMT: *office fédéral*) [...]

SRC Nur dieses Amt war in der Lage, [...]

SMT que ce *poste* tait dans la situation, [...]

Machine Translation

Coreferenced German compounds

Die Originalauswertung wurde in den Zwischenmassstab 1:200 reduziert, worauf das Bundesamt (SMT: *office fédéral*) [...]

coreference

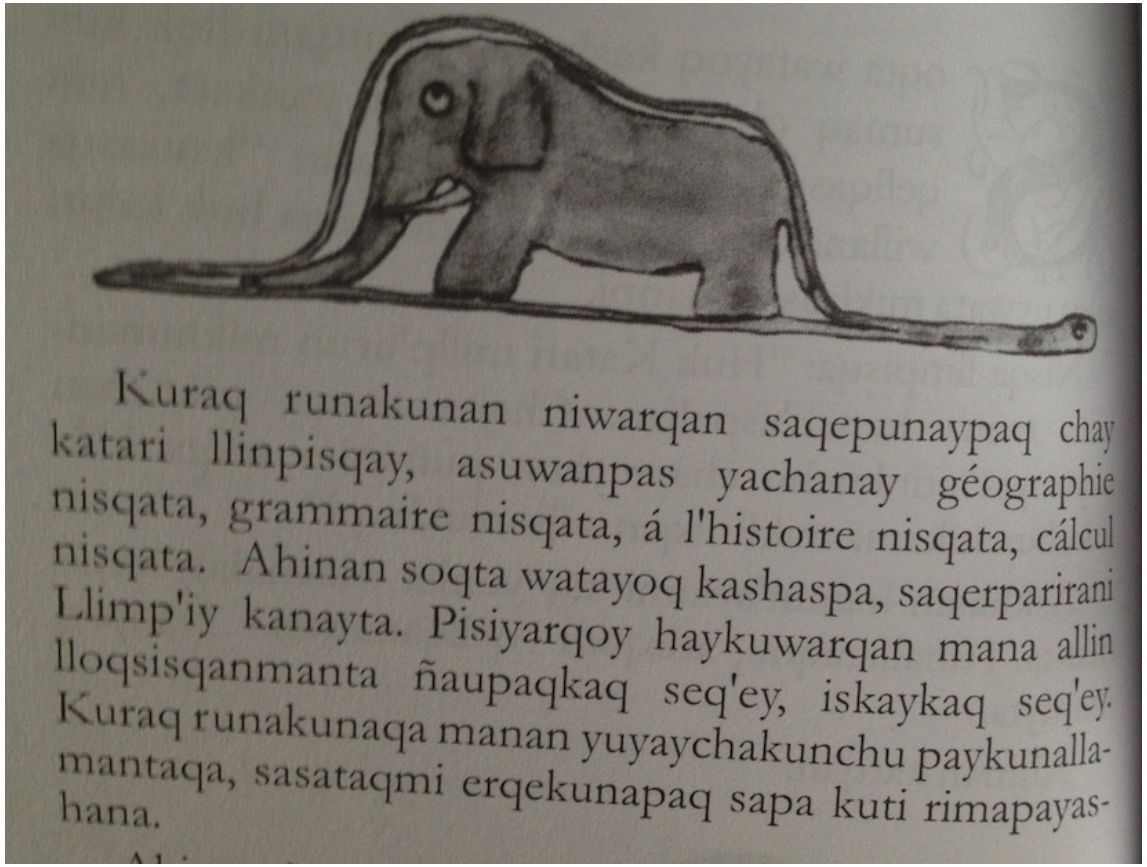
SRC Nur dieses Amt war in der Lage, [...]

SMT que ce *poste* tait dans la situation, [...]

que ce office tait dans la situation, [...]

Machine Translation

Low-Resource Languages



Kuraq runakunan niwarqan saqepunaypaq chay katari llinpisqay, asuwanpas yachanay géographie nisqata, grammaire nisqata, á l'histoire nisqata, cálcul nisqata. Ahinan soqta watayoq kashaspa, saqerparirani Llimp'iy kanayta. Pisiyarqoy haykuwarqan mana allin lloqsisqanmanta ñaupaqkaq seq'ey, iskaykaq seq'ey. Kuraq runakunaqa manan yuyaychakunchu paykunallamantaqa, sasataqmi erqekunapaq sapa kuti rimapayashana.

[que] tarpuysirichikunayawasqaykichikmantallañpunichá

[eng] “certainly, though, since you tried to inspire in me the wish to just help you sow”

Common Concepts

Annotation

Common Concepts

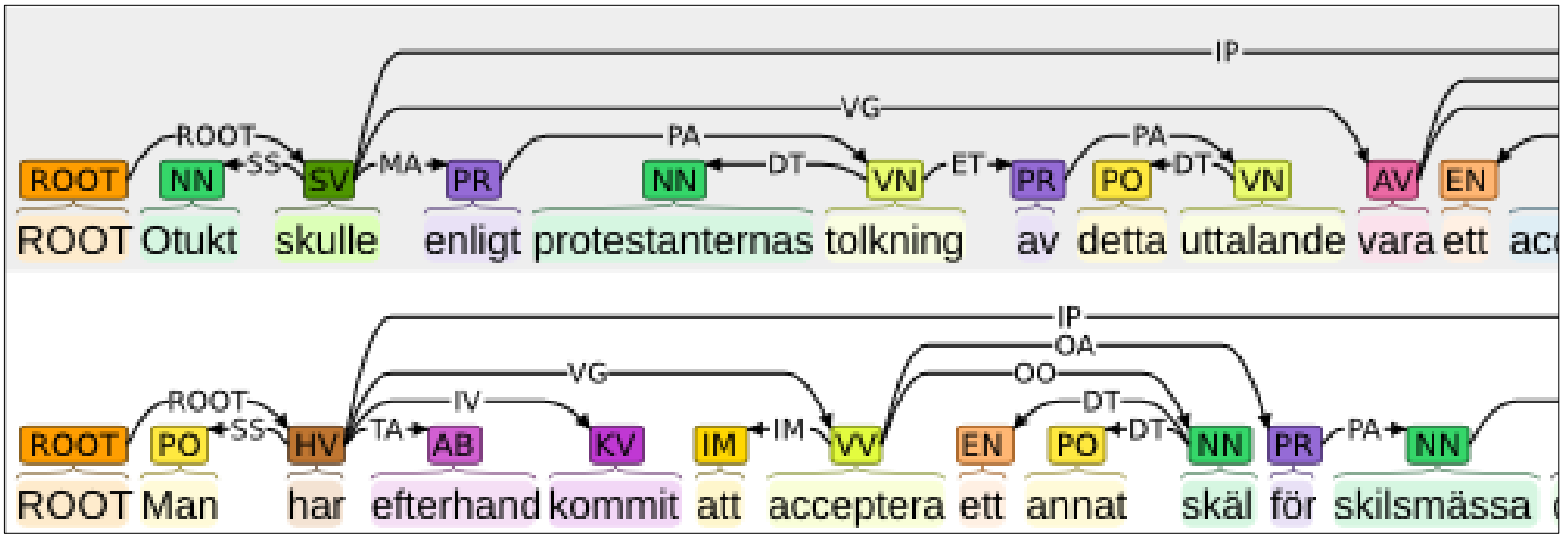
Annotation

- Part of spech (Noun, Verb, Adjective, …)

Common Concepts

Annotation

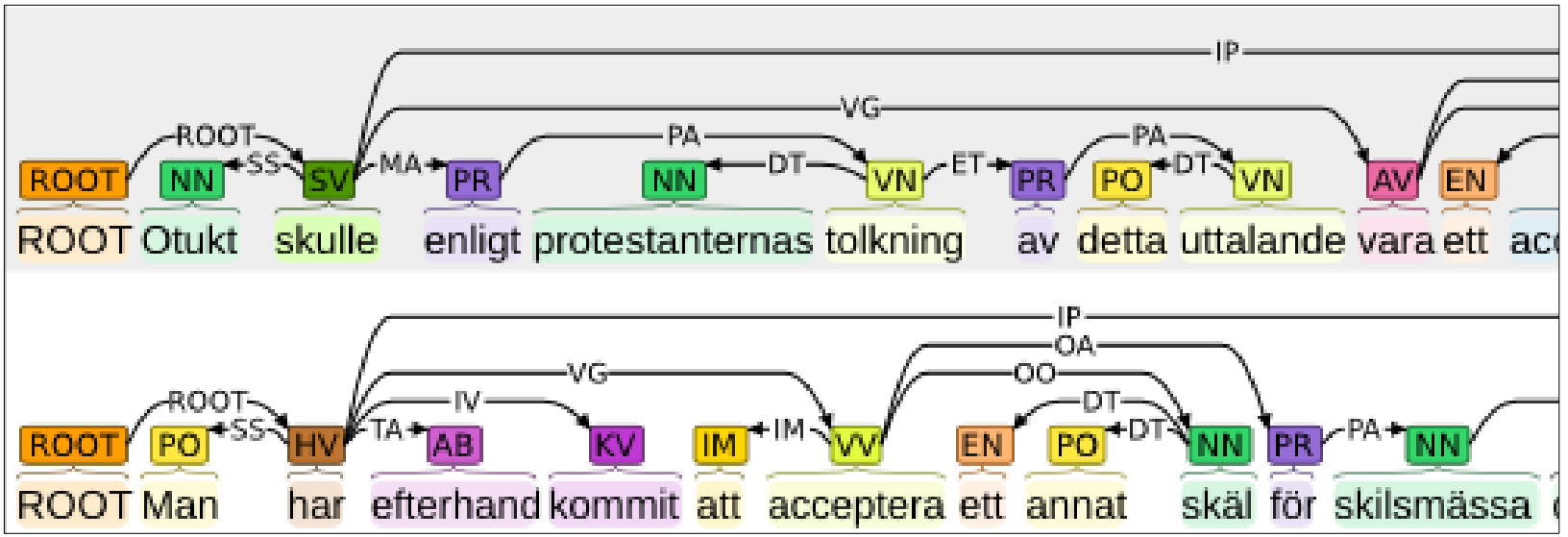
- Part of spech (Noun, Verb, Adjective, …)
- Syntactic relations (Subject, Object, Attribute, ...)



Common Concepts

Annotation

- Part of spech (Noun, Verb, Adjective, …)
- Syntactic relations (Subject, Object, Attribute, ...)
- Coreference (the teacher → he/she, ...)



Common Concepts

Alignment

Common Concepts

Alignment

- Sentence alignment

[eng] Mrs Banotti, if I may say so, you have anticipated the matter by about five minutes. We are just about to begin a debate on the matter with Mr Liikanen.

[ita] Onorevole Banotti, se permette, lei ha anticipato di qualche minuto la discussione che avrà luogo sull'argomento con il Commissario Liikanen.

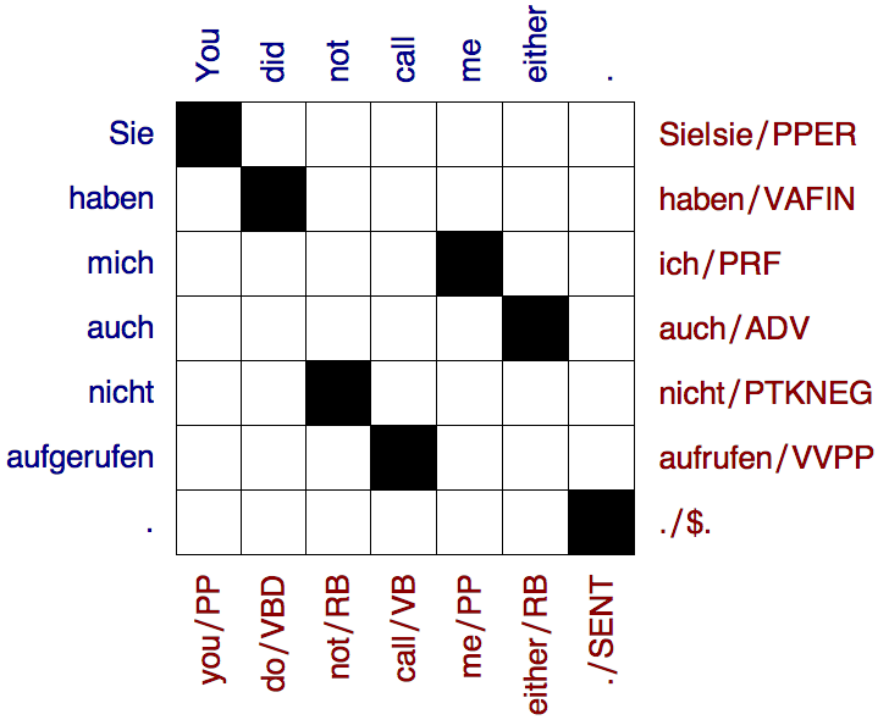
Common Concepts

Alignment

- Sentence alignment
- Word alignment (tokens)

[eng] Mrs Banotti, if I may say so, you have anticipated the matter by about five minutes. We are just about to begin a debate on the matter with Mr Liikanen.

[ita] Onorevole Banotti, se permette, lei ha anticipato di qualche minuto la discussione che avrà luogo sull'argomento con il Commissario Liikanen.



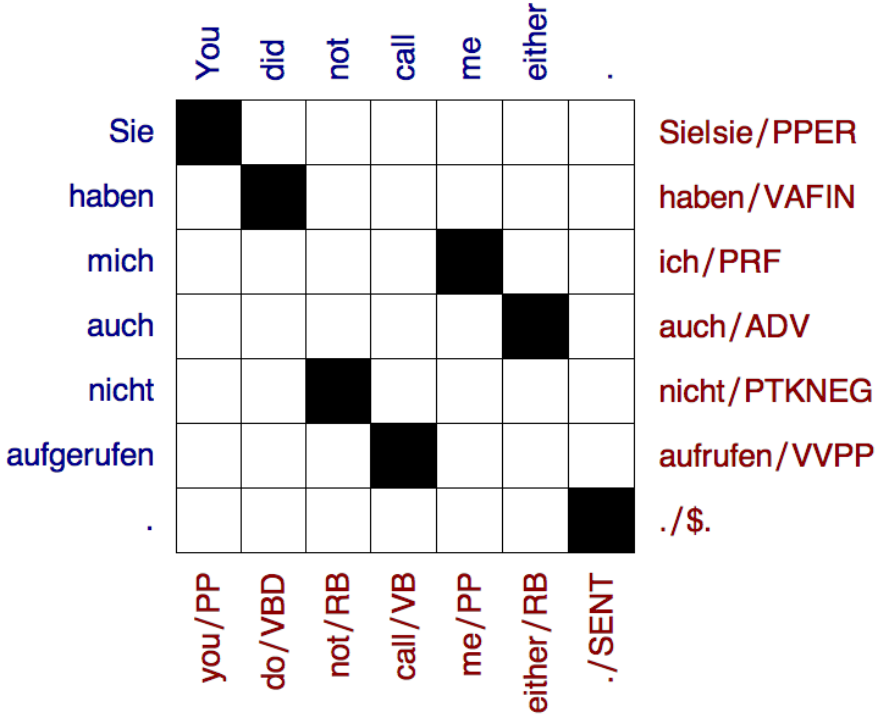
Common Concepts

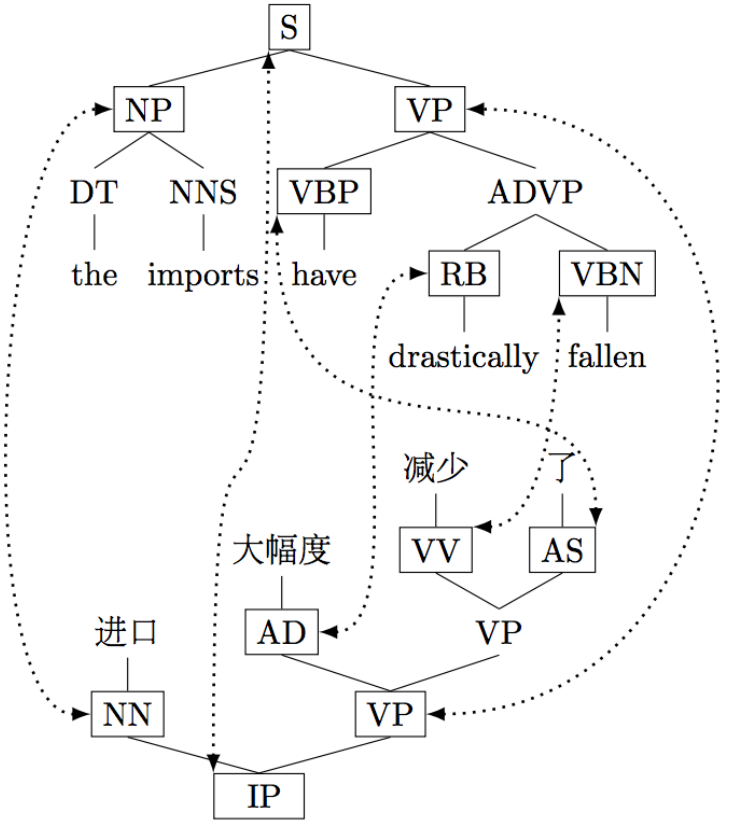
Alignment

- Sentence alignment
- Word alignment (tokens)
- Tree alignment (syntax trees)

[eng] Mrs Banotti, if I may say so, you have anticipated the matter by about five minutes. We are just about to begin a debate on the matter with Mr Liikanen.

[ita] Onorevole Banotti, se permette, lei ha anticipato di qualche minuto la discussione che avrà luogo sull'argomento con il Commissario Liikanen.





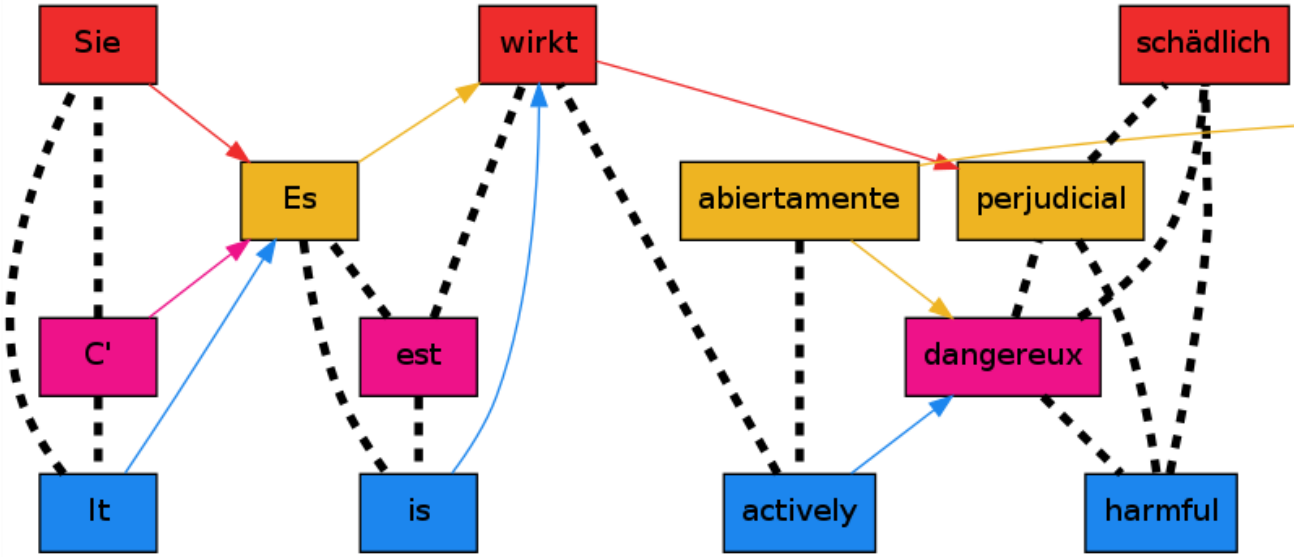
Large Text Corpora

Building a large multi-parallel corpus for linguistic studies

Large Text Corpora

Building a large multi-parallel corpus for linguistic studies

- More than two languages

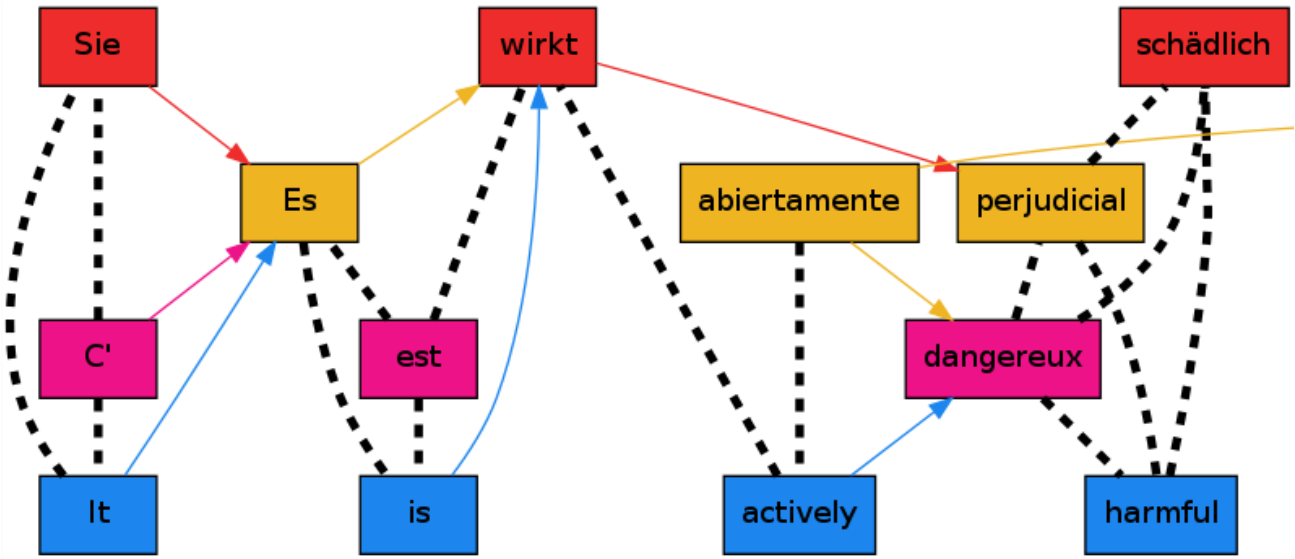


Large Text Corpora

Building a large multi-parallel corpus for linguistic studies

- More than two languages

  → Better models & error correction



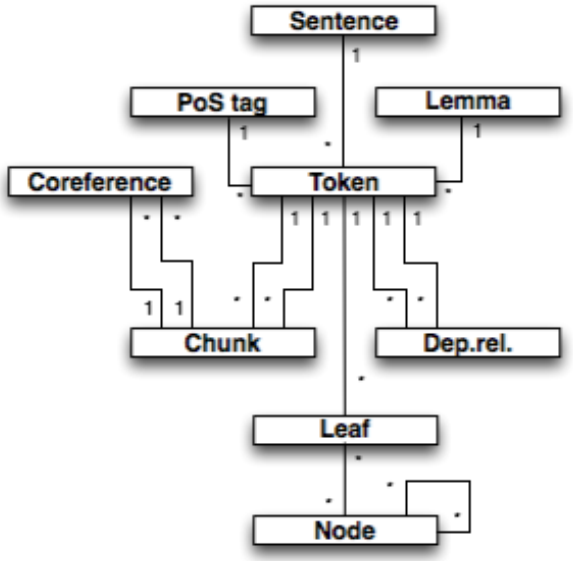
Large Text Corpora

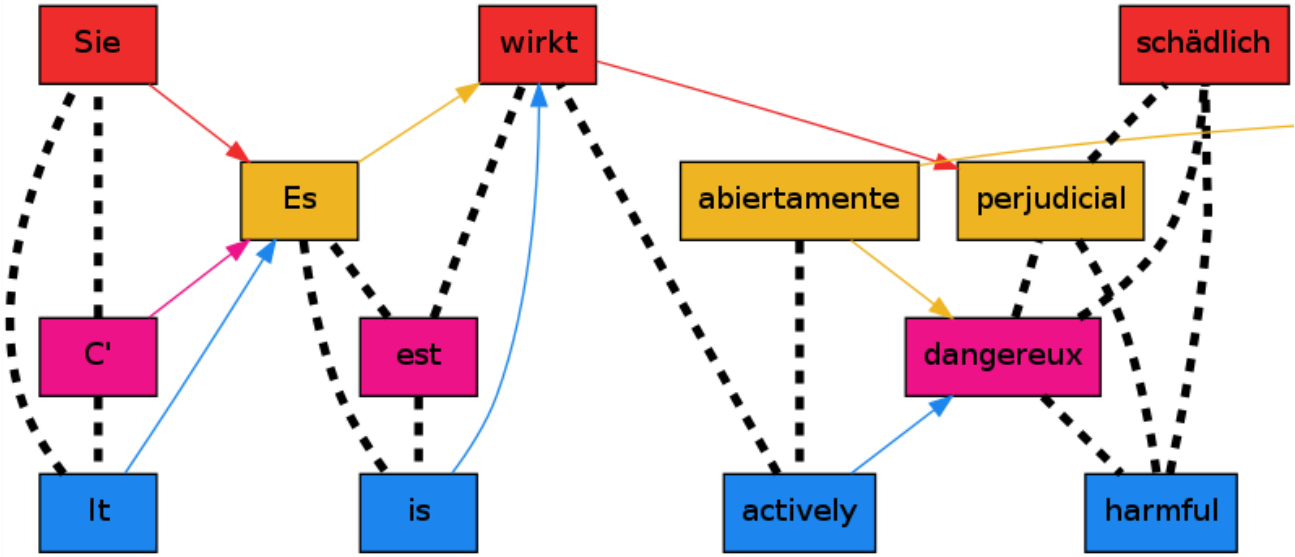
Building a large multi-parallel corpus for linguistic studies

- More than two languages

  → Better models & error correction

- Several layers of annotation and alignment





Large Text Corpora

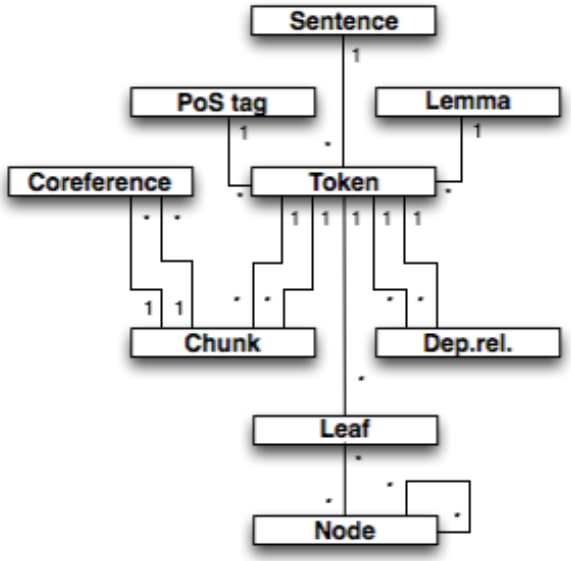
Building a large multi-parallel corpus for linguistic studies

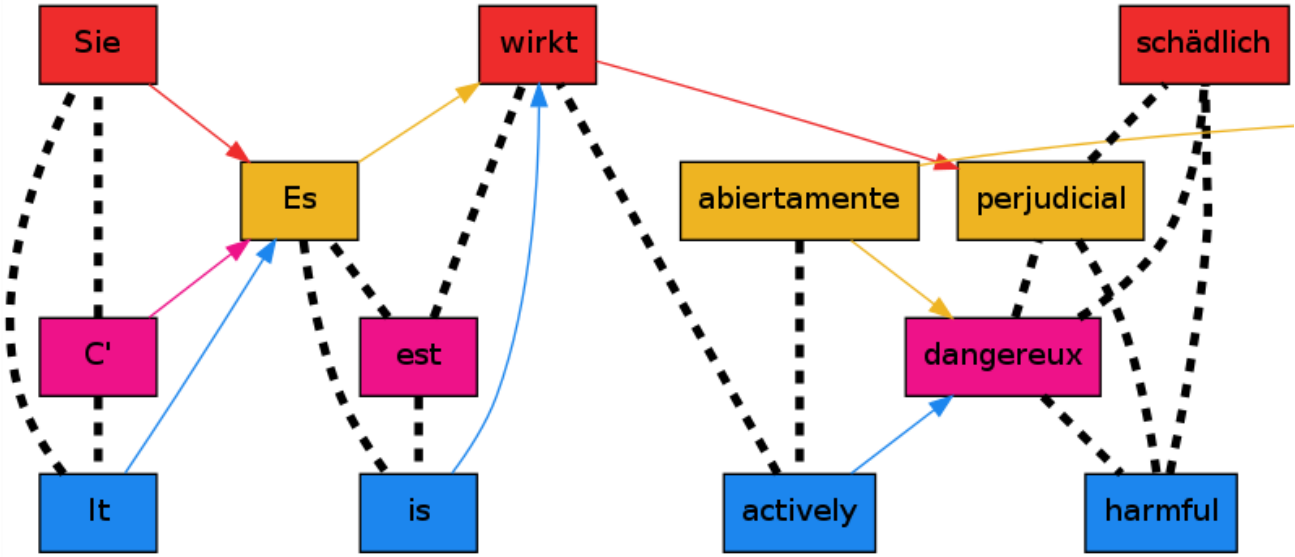
- More than two languages

  → Better models & error correction

- Several layers of annotation and alignment

  → Answering complex linguistic questions





Large Text Corpora Databases

```
WITH RECURSIVE l4 AS
(
        WITH l3 AS
        (
                WITH RECURSIVE l2 AS
                (
                        WITH l1 AS -- lno, max_lno, token_id, tnop, tlenp, req, avb, ava, reqb, reqa
                        (
                                SELECT lno, MAX(lno) OVER () max_lno, text_id token_id, tnop, tlenp, nterm, europarl2.array_agg_notnull(CASE WHEN nterm THEN t
                                        COUNT(*) OVER (PARTITION BY text_id ORDER BY tno ASC ROWS UNBOUNDED PRECEDING) - 1 avb,
                                        COUNT(*) OVER (PARTITION BY text_id ORDER BY tno DESC ROWS UNBOUNDED PRECEDING) - 1 ava,
                                        SUM(nterm::INT) OVER (PARTITION BY text_id ORDER BY tno ASC ROWS UNBOUNDED PRECEDING) - nterm::INT reqb,
                                        SUM(nterm::INT) OVER (PARTITION BY text_id ORDER BY tno DESC ROWS UNBOUNDED PRECEDING) - nterm::INT reqa
                                FROM
                                (
                                        SELECT *, (europarl2.get_term3(text_id)).*
                                        FROM
                                        (
                                                SELECT ROW_NUMBER() OVER (ORDER BY language_id ASC) lno, text_id, language_id
                                                FROM europarl2.TEXT
                                                WHERE turn_id = 19415
                                                AND language_id IN (1550,1834,1948,5996)
                                        ) x
                                ) y
                        )
                        SELECT l1.lno, l1.max_lno, ARRAY[l1.token_id] token_ids, l1.req, l1.tnop tnopl, l1.tnop tnoph, l1.tlenp tlenpl, l1.tlenp tlenp
                        FROM l1
                        WHERE lno = 1
                        UNION ALL
                        SELECT l1.lno, l1.max_lno, l2.token_ids||l1.token_id, l1.req, LEAST(l1.tnop,l2.tnopl), GREATEST(l1.tnop,l2.tnoph),
                                LEAST(l1.tlenp,l2.tlenpl), GREATEST(l1.tlenp,l2.tlenph),
                                GREATEST(l2.avb,l1.avb), GREATEST(l2.ava,l1.ava), GREATEST(l2.reqb,l1.reqb), GREATEST(l2.reqa,l1.reqa)
                        FROM l1, l2
                        WHERE l1.lno = (l2.lno+1)
                        AND l2.avb >= l1.reqb AND l2.ava >= l1.reqa AND l2.reqb <= l1.avb AND l2.reqa <= l1.ava
                        AND GREATEST(l1.tnop,l2.tnoph) - LEAST(l1.tnop,l2.tnopl) < 0.2
                )
                SELECT ROW_NUMBER() OVER () l2_no, token_ids, req, tnoph-tnopl delta, tnoph-tnopl tnopd, tlenph-tlenpl tlenpd, tnopl, tnoph
                FROM l2
                WHERE lno = max_lno
        )
        SELECT l3.l2_no, l3.token_ids token_ids, europarl2.array_subtraction(req,l3.token_ids) rest, ARRAY[token_ids] groups, 1-SQRT(delta) score, tn
        FROM l3
        UNION ALL
        SELECT l3.l2_no, ARRAY(SELECT UNNEST(array_cat(l4.token_ids,l3.token_ids))), europarl2.array_subtraction(l4.rest,l3.token_ids),
                array_cat(l4.groups,l3.token_ids), l4.score + (1-SQRT(delta)),
                LEAST(l3.tnopl,l4.tnopl) tnopl, GREATEST(l3.tnoph,l4.tnoph) tnoph
        FROM l3, l4
        WHERE NOT l3.token_ids && l4.token_ids
        AND l3.l2_no > l4.l2_no
        AND (l3.tnopl > l4.tnoph OR l3.tnoph < l4.tnopl)
)
SELECT groups, score
FROM l4
WHERE array_upper(rest,1) IS NULL
```

Large Text Corpora

Databases

- Annotations and alignments are relations,

```
WITH RECURSIVE l4 AS
(
	WITH l3 AS
	(
		WITH RECURSIVE l2 AS
		(
			WITH l1 AS -- lno, max_lno, token_id, tnop, tlenp, req, avb, ava, reqb, reqa
			(
				SELECT lno, MAX(lno) OVER () max_lno, token_id, tnop, tlenp, nterm, europarl2.array_agg_notnull(CASE WHEN nterm THEN t
					COUNT(*) OVER (PARTITION BY text_id ORDER BY tno ASC ROWS UNBOUNDED PRECEDING) - 1 avb,
					COUNT(*) OVER (PARTITION BY text_id ORDER BY tno DESC ROWS UNBOUNDED PRECEDING) - 1 ava,
					SUM(nterm::INT) OVER (PARTITION BY text_id ORDER BY tno ASC ROWS UNBOUNDED PRECEDING) - nterm::INT reqb,
					SUM(nterm::INT) OVER (PARTITION BY text_id ORDER BY tno DESC ROWS UNBOUNDED PRECEDING) - nterm::INT reqa
				FROM
					SELECT *, (europarl2.get_term3(text_id)).*
					FROM
					(
						SELECT ROW_NUMBER() OVER (ORDER BY language_id ASC) lno, text_id, language_id
						FROM europarl2.TEXT
						WHERE turn_id = 19415
						AND language_id IN (1550,1834,1948,5996)
					) x
				) y
			)
			SELECT l1.lno, l1.max_lno, ARRAY[l1.token_id] token_ids, l1.req, l1.tnop tnopl, l1.tnop tnoph, l1.tlenp tlenpl, l1.tlenp tlenp
			FROM l1
			WHERE lno = 1
			UNION ALL
			SELECT l1.lno, l1.max_lno, l2.token_ids||l1.token_id, l1.req, LEAST(l1.tnop,l2.tnopl), GREATEST(l1.tnop,l2.tnoph),
				LEAST(l1.tlenp,l2.tlenpl), GREATEST(l1.tlenp,l2.tlenph),
				GREATEST(l2.avb,l1.avb), GREATEST(l2.ava,l1.ava), GREATEST(l2.reqb,l1.reqb), GREATEST(l2.reqa,l1.reqa)
			FROM l1, l2
			WHERE l1.lno = (l2.lno+1)
			AND l2.avb >= l1.reqb AND l2.ava >= l1.reqa AND l2.reqb <= l1.avb AND l2.reqa <= l1.ava
			AND GREATEST(l1.tnop,l2.tnoph) - LEAST(l1.tnop,l2.tnopl) < 0.2
		)
		SELECT ROW_NUMBER() OVER () l2_no, token_ids, req, tnoph-tnopl delta, tnoph-tnopl tnopd, tlenph-tlenpl tlenpd, tnopl, tnoph
		FROM l2
		WHERE lno = max_lno
	)
	SELECT l3.l2_no, l3.token_ids token_ids, europarl2.array_subtraction(req,l3.token_ids) rest, ARRAY[token_ids] groups, 1-SQRT(delta) score, tno
	FROM l3
	UNION ALL
	SELECT l3.l2_no, ARRAY(SELECT UNNEST(array_cat(l4.token_ids,l3.token_ids))), europarl2.array_subtraction(l4.rest,l3.token_ids),
		array_cat(l4.groups,l3.token_ids), l4.score + (1-SQRT(delta)),
		LEAST(l3.tnopl,l4.tnopl) tnopl, GREATEST(l3.tnoph,l4.tnoph) tnoph
	FROM l3, l4
	WHERE NOT l3.token_ids && l4.token_ids
	AND l3.l2_no > l4.l2_no
	AND (l3.tnopl > l4.tnoph OR l3.tnoph < l4.tnopl)
)
SELECT groups, score
FROM l4
WHERE array_upper(rest,1) IS NULL
```

Large Text Corpora

Databases

- Annotations and alignments are relations,
- We have large corpora,

```
WITH RECURSIVE l4 AS
(
        WITH l3 AS
        (
                WITH RECURSIVE l2 AS
                (
                        WITH l1 AS -- lno, max_lno, token_id, tnop, tlenp, req, avb, ava, reqb, reqa
                        (
                                SELECT lno, MAX(lno) OVER () max_lno, token_id, tnop, tlenp, nterm, europarl2.array_agg_notnull(CASE WHEN nterm THEN t
                                        COUNT(*) OVER (PARTITION BY text_id ORDER BY tno ASC ROWS UNBOUNDED PRECEDING) - 1 avb,
                                        COUNT(*) OVER (PARTITION BY text_id ORDER BY tno DESC ROWS UNBOUNDED PRECEDING) - 1 ava,
                                        SUM(nterm::INT) OVER (PARTITION BY text_id ORDER BY tno ASC ROWS UNBOUNDED PRECEDING) - nterm::INT reqb,
                                        SUM(nterm::INT) OVER (PARTITION BY text_id ORDER BY tno DESC ROWS UNBOUNDED PRECEDING) - nterm::INT reqa
                                FROM
                                        SELECT *, (europarl2.get_term3(text_id)).*
                                        FROM
                                        (
                                                SELECT ROW_NUMBER() OVER (ORDER BY language_id ASC) lno, text_id, language_id
                                                FROM TEXT
                                                WHERE text_id = 2415
                                                AND language_id IN (1550,1834,1948,5996)
                                        ) x
                                ) y
                        )
                        SELECT l1.lno, l1.max_lno, ARRAY[l1.token_id] token_ids, l1.req, l1.tnop tnopl, l1.tnop tnoph, l1.tlenp tlenpl, l1.tlenp tlenp
                        FROM l1
                        WHERE lno = 1
                        UNION ALL
                        SELECT l1.lno, l1.max_lno, l2.token_ids||l1.token_id, l1.req, LEAST(l1.tnop,l2.tnopl), GREATEST(l1.tnop,l2.tnoph),
                                LEAST(l1.tlenp,l2.tlenpl), GREATEST(l1.tlenp,l2.tlenph),
                                GREATEST(l2.avb,l1.avb), GREATEST(l2.ava,l1.ava), GREATEST(l2.reqb,l1.reqb), GREATEST(l2.reqa,l1.reqa)
                        FROM l1, l2
                        WHERE l1.lno = (l2.lno+1)
                        AND l2.avb >= l1.reqb AND l2.ava >= l1.reqa AND l2.reqb <= l1.avb AND l2.reqa <= l1.ava
                        AND GREATEST(l1.tnop,l2.tnoph) - LEAST(l1.tnop,l2.tnopl) < 0.2
                )
                SELECT ROW_NUMBER() OVER () l2_no, token_ids, req, tnoph-tnopl delta, tnoph-tnopl tnopd, tlenph-tlenpl tlenpd, tnopl, tnoph
                FROM l2
                WHERE lno = max_lno
        )
        SELECT l3.l2_no, l3.token_ids token_ids, europarl2.array_subtraction(req,l3.token_ids) rest, ARRAY[token_ids] groups, 1-SQRT(delta) score, tno
        FROM l3
        UNION ALL
        SELECT l3.l2_no, ARRAY(SELECT UNNEST(array_cat(l4.token_ids,l3.token_ids))), europarl2.array_subtraction(l4.rest,l3.token_ids),
                array_cat(l4.groups,l3.token_ids), l4.score + (1-SQRT(delta)),
                LEAST(l3.tnopl,l4.tnopl) tnopl, GREATEST(l3.tnoph,l4.tnoph) tnoph
        FROM l3, l4
        WHERE NOT l3.token_ids && l4.token_ids
        AND l3.l2_no > l4.l2_no
        AND (l3.tnopl > l4.tnoph OR l3.tnoph < l4.tnopl)
)
SELECT groups, score
FROM l4
WHERE array_upper(rest,1) IS NULL
```

Large Text Corpora

Databases

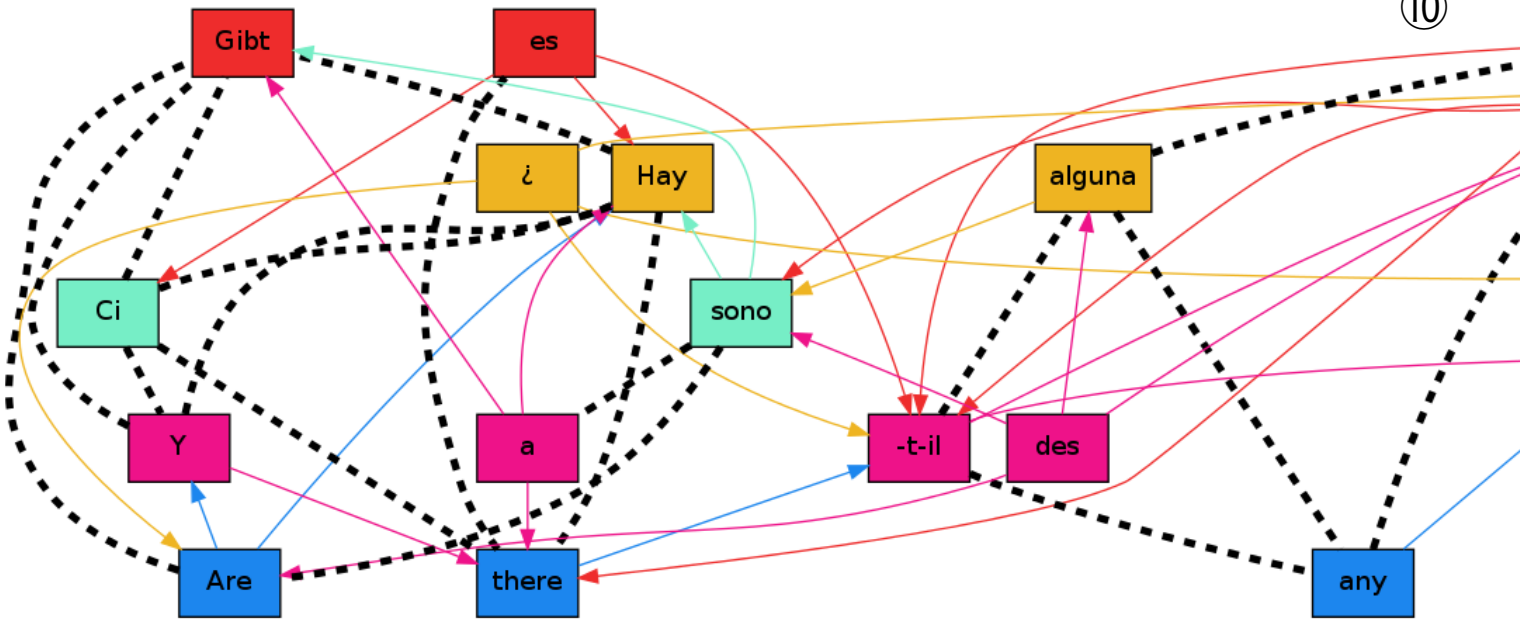
- Annotations and alignments are relations,
- We have large corpora,
- … and want to execute complex queries.

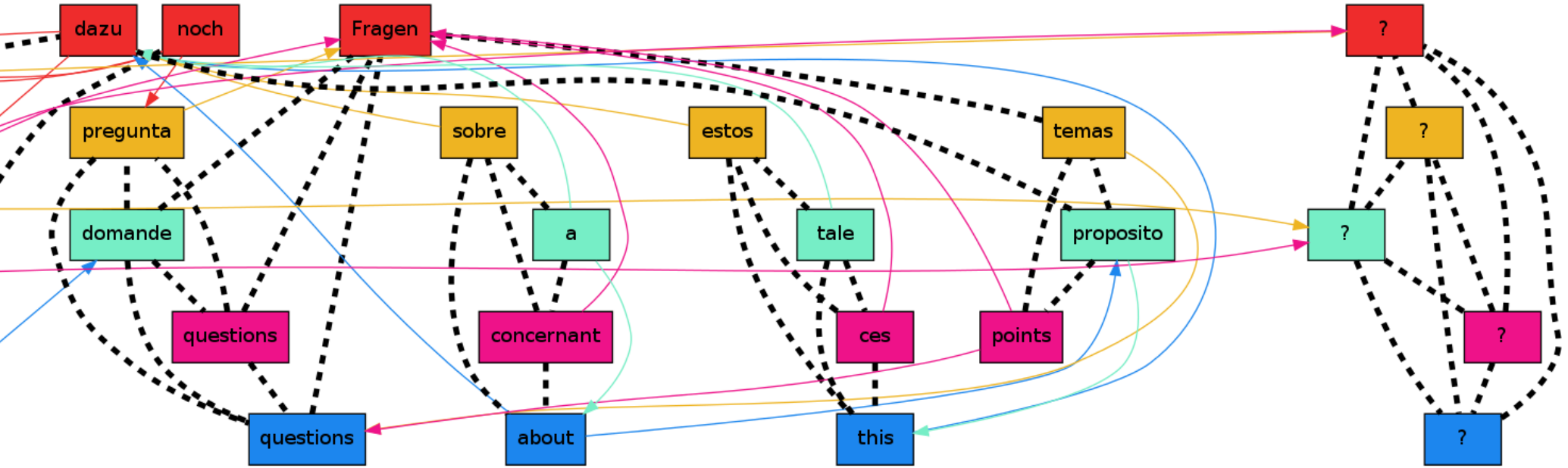
Large Text Corpora

Databases

- Annotations and alignments are relations,
- We have large corpora,
- … and want to execute complex queries.

→ Relational Databases


Gibt
es
¿
Hay
alguna
Ci
sono
Y
a
-t-il
des
Are
there
any


dazu
noch
Fragen
?
pregunta
sobre
estos
temas
?
domande
a
tale
proposito
?
questions
concernant
ces
points
?
questions
about
this
?