Language technology

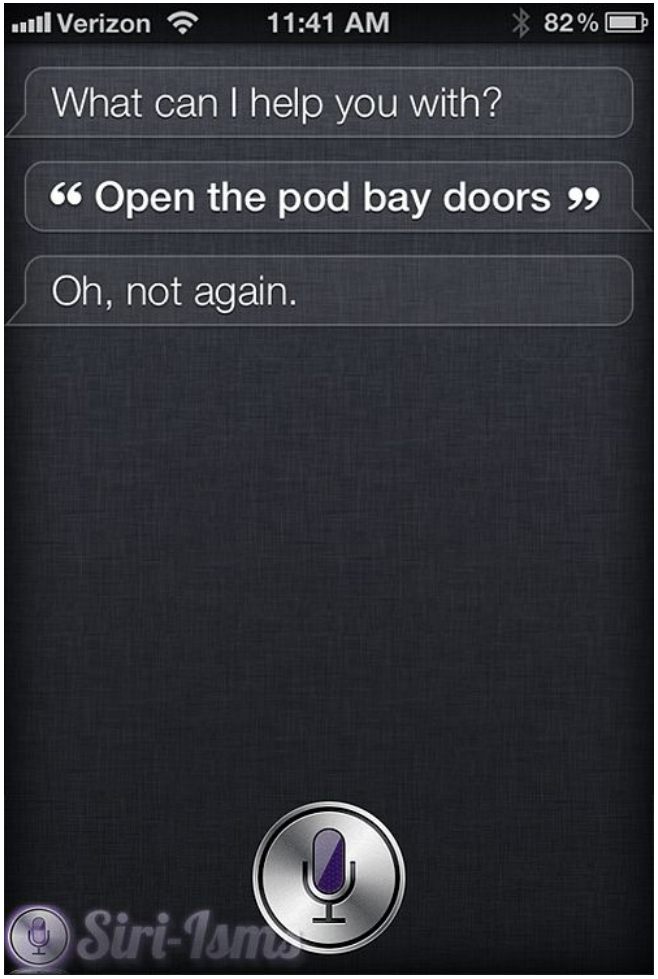
Research, development, participation

November 5, 2015
Sissejuhatus informaatikasse
Mark Fišel, keeletehnoloogia uurimisrühm

Outline

- Dialogue systems
- Machine translation

Dialogue systems



Simple dialogue systems

Kasutaja: mitu arvutiklassi on?

Süsteem: mat-inf majas on neli arvutiklassi, kus on peal Windowsi ning Linuxi operatsioonisüsteemid

Kasutaja: kas wifi on majas tasuta?

Süsteem: Tartu Ülikool pakub tasuta traadita interneti teenust enamikes õppehoonetes

Kasutaja: mitu aastat tuleb ülikoolis käia, et saada baka?

Süsteem: bakalaureuseõpe on kõrghariduse esimene aste, mille nominaalkestus on kolm aastat

Kasutaja: kas IT-sektoris on tööd kõvasti?

More interesting dialogue systems

Kasutaja: tahaks laupäeval teatrisse minna

Süsteem: mis teatris te soovite etendust vaadata?

Kasutaja: mingi tallinna teater näiteks.

Süsteem: meie andmebaasis on etendusi järgmistest Tallinna teatritest: Tallinna Linnateater, Teater No99, Salme Kultuurikeskus ja Vene Teater

Kasutaja: mis siis linnateatris laupäeval toimub?

Süsteem: Tallinna Linnateatris mängitakse 7. novembril etendusi: “Melujanu”, “Igatsus”

Kasutaja: läheks siis “igatsust” vaatama

Süsteem: etendust “igatsus” mängitakse laupäeval, 7. novembril kell 18:00

How it works (simple)?

- define keywords/patterns
  - e.g. “… wifi …”
  - e.g. “… kaua/aasta … baka/bakalaureuse …”

How it works (simple)?

- define keywords/patterns
  - e.g. “… wifi …”
  - e.g. “… kaua/aasta … baka/bakalaureuse …”
- define reactions to these patterns
  - e.g. “TÜ pakub tasuta traadita interneti teenust enamikes õppehoonetes”
  - e.g. “bakalaureuseõpe on kõrghariduse esimene aste, mille nominaalkestus on kolm aastat”

How it works (simple)?

- define keywords/patterns
  - e.g. “… wifi …”
  - e.g. “… kaua/aasta … baka/bakalaureuse …”
- define reactions to these patterns
  - e.g. “TÜ pakub tasuta traadita interneti teenust enamikes õppehoonetes”
  - e.g. “bakalaureuseõpe on kõrghariduse esimene aste, mille nominaalkestus on kolm aastat”
- ...profit!

How it works (more advanced)?

- define frames of information
  - e.g. *⟨event, date, time, place⟩*

How it works (more advanced)?

- define frames of information
  - e.g. ⟨*event, date, time, place*⟩
- match the patterns to frame entries
  - e.g. “...homme/reedel/…” → extract date, fill frame

How it works (more advanced)?

- define frames of information
  - e.g. ⟨*event, date, time, place*⟩
- match the patterns to frame entries
  - e.g. “...homme/reedel/…” → extract date, fill frame
- when missing a piece of info, ask for it
  - e.g. “mis teatris te soovite etendust vaadata?”

How it works (more advanced)?

- define frames of information
  - e.g. ⟨*event, date, time, place*⟩
- match the patterns to frame entries
  - e.g. “...homme/reedel/…” → extract date, fill frame
- when missing a piece of info, ask for it
  - e.g. “mis teatris te soovite etendust vaadata?”
- when the frame has all the necessary info, make a DB query and shape an answer
  - find all answers matching the date, time and place
  - shape report: “etendust X mängitakse Y kell Z”

What makes it ~~hard~~ more interesting?

- words are in different morphological forms
  - teater / teatri / teatris / …

What makes it ~~hard~~ more interesting?

- words are in different morphological forms
  - teater / teatri / teatris / …
- language is ambiguous
  - time flies like an arrow = *ajakärbestele meeldib nool??*

What makes it ~~hard~~ more interesting?

- words are in different morphological forms
  - teater / teatri / teatris / …
- language is ambiguous
  - time flies like an arrow =
    ajakärbestele meeldib nool??
  - otsi vead üles =
    (you) are pulling the ends up??

What makes it ~~hard~~ more interesting?

- words are in different morphological forms
  - teater / teatri / teatris / …
- language is ambiguous
  - time flies like an arrow =
    ajakärbestele meeldib nool??
  - otsi vead üles =
    (you) are pulling the ends up??
- many versions of the same meaning possible
  - homme / 6. novembril / reedel / …

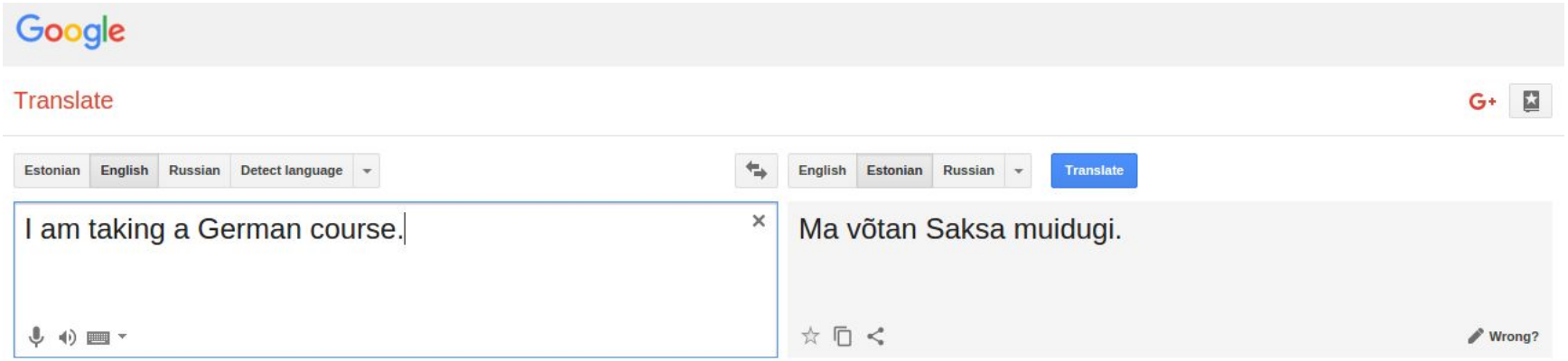
Is that it?

Is that it?

No!

- dialogue act recognition
- dialogues on broader topics
- automatically adaptive systems
- …

Machine translation



Machine translation @UT

Two approaches to machine translation

1. Let us manually describe the bilingual dictionary, reordering rules, etc.

For example, En→Fr:

Dictionary:

words:

- house → maison
- car → voiture

phrases:

- of course → bien sur

…

Reordering rules:

- Adjective_{EN} Noun_{EN} → Noun_{FR} Adjective_{FR}
  - e.g. “red house” = “maison rouge”
- …

Two approaches to machine translation

1. Let us manually describe the bilingual dictionary, reordering rules, etc.

For example, En→Fr:

Dictionary:

words:

- house → maison
- car → voiture

phrases:

- of course → bien sur

NB! the → le? la?

…

Reordering rules:

- Adjective_{EN} Noun_{EN} → Noun_{FR} Adjective_{FR}
  - e.g. “red house” = “maison rouge”
- …

Two approaches to machine translation

1. Let us manually describe the bilingual dictionary, reordering rules, etc.

For example, En→Fr:

Dictionary:

words:

- house → maison
- car → voiture

phrases:

- of course → bien sur

NB! the → le? la?

…

Reordering rules:

- Adjective_{EN} Noun_{EN} → Noun_{FR} Adjective_{FR}
  - e.g. “red house” = “maison rouge”
- NB! “the little corporal” → “le caporal petit”??
- …

Two approaches to machine translation

1. Let us manually describe the bilingual dictionary, reordering rules, etc.

   For example, En→Fr:

   a. long and boring

   Dictionary:
   words:
   - house → maison
   - car → voiture

   phrases:
   - of course → bien sur

   NB! the → le? la?

   …

   b. must write very complicated rules to translate reasonable texts

   Reordering rules:
   - $\text{Adjective}_{EN}\ \text{Noun}_{EN} \rightarrow \text{Noun}_{FR}\ \text{Adjective}_{FR}$
   - e.g. "red house" = "maison rouge"
   - NB! "little cousin" → "cousin petit"??
   - …

   c. new language pair = start from the very beginning

Centauri-Arcturan Translation

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

Centauri-Arcturan Translation

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok **ghirok** .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = ?

Centauri-Arcturan Translation

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok **ghirok** .

3b. totat dat arrat vat **hilat** .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = ?

Centauri-Arcturan Translation

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok **ghirok** .

3b. totat dat arrat vat **hilat** .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

Centauri-Arcturan Translation

1a. **ok-voon** ororok sprok .

1b. **at-voon** bichat dat .

2a. **ok-drubel ok-voon** anok plok sprok .

2b. **at-drubel at-voon** pippat rrat dat .

3a. erok sprok izok hihok **ghirok** .

3b. totat dat arrat vat **hilat** .

4a. **ok-voon** anok drok brok jok .

4b. **at-voon** krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok **ok-yurp** .

9b. totat nnat quat oloat **at-yurp** .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok **zanzanok** .

11b. wat nnat arrat mat **zanzanat** .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = ?
ok-drubel = ?
ok-yurp = ?
zanzanok = ?

Centauri-Arcturan Translation

1a. **ok-voon** ororok sprok .

1b. **at-voon** bichat dat .

2a. **ok-drubel ok-voon** anok plok sprok .

2b. **at-drubel at-voon** pippat rrat dat .

3a. erok sprok izok hihok **ghirok** .

3b. totat dat arrat vat **hilat** .

4a. **ok-voon** anok drok brok jok .

4b. **at-voon** krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok **ok-yurp** .

9b. totat nnat quat oloat **at-yurp** .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok **zanzanok** .

11b. wat nnat arrat mat **zanzanat** .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = ?
ok-drubel = ?
ok-yurp = ?
zanzanok = ?

Centauri-Arcturan Translation

1a. **ok-voon** ororok sprok .

1b. **at-voon** bichat dat .

2a. **ok-drubel ok-voon** anok plok sprok .

2b. **at-drubel at-voon** pippat rrat dat .

3a. erok sprok izok hihok **ghirok** .

3b. totat dat arrat vat **hilat** .

4a. **ok-voon** anok drok brok jok .

4b. **at-voon** krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok **ok-yurp** .

9b. totat nnat quat oloat **at-yurp** .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok **zanzanok** .

11b. wat nnat arrat mat **zanzanat** .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

Centauri-Arcturan Translation

1a. **ok-voon** ororok **sprok** .

1b. **at-voon** bichat **dat** .

2a. **ok-drubel** **ok-voon** anok plok **sprok** .

2b. **at-drubel** **at-voon** pippat rrat **dat** .

3a. erok **sprok** izok hihok **ghirok** .

3b. totat **dat** arrat vat **hilat** .

4a. **ok-voon** anok drok brok jok .

4b. **at-voon** krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok farok ororok lalok **sprok** izok enemok .

7b. wat jjat bichat wat **dat** vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok **ok-yurp** .

9b. totat nnat quat oloat **at-yurp** .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok **zanzanok** .

11b. wat nnat arrat mat **zanzanat** .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon** anok plok **sprok** .

2b. **at-drubel at-voon** pippat rrat **dat** .

3a. erok **sprok** izok hihok **ghirok** .

3b. totat **dat** arrat vat **hilat** .

4a. **ok-voon** anok drok brok jok .

4b. **at-voon** krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok farok **ororok** lalok **sprok** izok enemok .

7b. wat jjat **bichat** wat **dat** vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok **ok-yurp** .

9b. totat nnat quat oloat **at-yurp** .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok **zanzanok** .

11b. wat nnat arrat mat **zanzanat** .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok** plok **sprok** .

2b. **at-drubel at-voon pippat** rrat **dat** .

3a. erok **sprok** izok hihok **ghirok** .

3b. totat **dat** arrat vat **hilat** .

4a. **ok-voon anok** drok brok jok .

4b. **at-voon** krat **pippat** sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok farok **ororok** lalok **sprok** izok enemok .

7b. wat jjat **bichat** wat **dat** vat eneat .

8a. lalok brok **anok** plok nok .

8b. iat lat **pippat** rrat nnat .

9a. wiwok nok izok kantok **ok-yurp** .

9b. totat nnat quat oloat **at-yurp** .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok **zanzanok** .

11b. wat nnat arrat mat **zanzanat** .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok hihok **ghirok** .

3b. totat **dat** arrat vat **hilat** .

4a. **ok-voon anok** drok brok jok .

4b. **at-voon** krat **pippat** sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok farok **ororok** lalok **sprok** izok enemok .

7b. wat jjat **bichat** wat **dat** vat eneat .

8a. lalok brok **anok plok** nok .

8b. iat lat **pippat rrat** nnat .

9a. wiwok nok izok kantok **ok-yurp** .

9b. totat nnat quat oloat **at-yurp** .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok hihok yorok **zanzanok** .

11b. wat nnat arrat mat **zanzanat** .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok **hihok ghirok** .

3b. totat **dat arrat** vat **hilat** .

4a. **ok-voon anok** drok brok jok .

4b. **at-voon** krat **pippat** sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok farok **ororok** lalok **sprok** izok enemok .

7b. wat jjat **bichat** wat **dat** vat eneat .

8a. lalok brok **anok plok** nok .

8b. iat lat **pippat rrat** nnat .

9a. wiwok nok izok kantok **ok-yurp** .

9b. totat nnat quat oloat **at-yurp** .

10a. lalok mok nok yorok **ghirok** clok .

10b. wat nnat gat mat bat **hilat** .

11a. lalok nok crrrok **hihok** yorok **zanzanok** .

11b. wat nnat **arrat** mat **zanzanat** .

12a. lalok rarok nok izok **hihok** mok .

12b. wat nnat forat **arrat** vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok **hihok ghirok** .

3b. totat **dat arrat** vat **hilat** .

4a. **ok-voon anok** drok brok jok .

4b. **at-voon** krat **pippat** sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok farok **ororok** lalok **sprok** izok enemok .

7b. wat jjat **bichat** wat **dat** vat eneat .

8a. lalok brok **anok plok nok** .

8b. iat lat **pippat rrat nnat** .

9a. wiwok **nok** izok kantok **ok-yurp** .

9b. totat **nnat** quat oloat **at-yurp** .

10a. lalok mok **nok** yorok **ghirok** clok .

10b. wat **nnat** gat mat bat **hilat** .

11a. lalok **nok** crrrok **hihok** yorok **zanzanok** .

11b. wat **nnat arrat** mat **zanzanat** .

12a. lalok rarok **nok** izok **hihok** mok .

12b. wat **nnat** forat **arrat** vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok **hihok ghirok** .

3b. totat **dat arrat** vat **hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok farok **ororok** lalok **sprok** izok enemok .

7b. wat jjat **bichat** wat **dat** vat eneat .

8a. lalok **brok anok plok nok** .

8b. iat **lat pippat rrat nnat** .

9a. wiwok **nok** izok kantok **ok-yurp** .

9b. totat **nnat** quat oloat **at-yurp** .

10a. lalok mok **nok** yorok **ghirok** clok .

10b. wat **nnat** gat mat bat **hilat** .

11a. lalok **nok** crrrok **hihok** yorok **zanzanok** .

11b. wat **nnat arrat** mat **zanzanat** .

12a. lalok rarok **nok** izok **hihok** mok .

12b. wat **nnat** forat **arrat** vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok **hihok ghirok** .

3b. totat **dat arrat** vat **hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok** farok izok stok .

5b. **totat** jjat quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok farok **ororok** lalok **sprok** izok enemok .

7b. wat jjat **bichat** wat **dat** vat eneat .

8a. lalok **brok anok plok nok** .

8b. iat **lat pippat rrat nnat** .

9a. **wiwok nok** izok kantok **ok-yurp** .

9b. **totat nnat** quat oloat **at-yurp** .

10a. lalok mok **nok** yorok **ghirok** clok .

10b. wat **nnat** gat mat bat **hilat** .

11a. lalok **nok** crrrok **hihok** yorok **zanzanok** .

11b. wat **nnat arrat** mat **zanzanat** .

12a. lalok rarok **nok** izok **hihok** mok .

12b. wat **nnat** forat **arrat** vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok **hihok ghirok** .

3b. totat **dat arrat** vat **hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok farok** izok stok .

5b. **totat jjat** quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok **farok ororok** lalok **sprok** izok enemok .

7b. wat **jjat bichat** wat **dat** vat eneat .

8a. lalok **brok anok plok nok** .

8b. iat **lat pippat rrat nnat** .

9a. **wiwok nok** izok kantok **ok-yurp** .

9b. **totat nnat** quat oloat **at-yurp** .

10a. lalok mok **nok** yorok **ghirok** clok .

10b. wat **nnat** gat mat bat **hilat** .

11a. lalok **nok** crrrok **hihok** yorok **zanzanok** .

11b. wat **nnat arrat** mat **zanzanat** .

12a. lalok rarok **nok** izok **hihok** mok .

12b. wat **nnat** forat **arrat** vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat
farok = jjat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok **hihok ghirok** .

3b. totat **dat arrat** vat **hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok farok** izok stok .

5b. **totat jjat** quat cat .

6a. lalok **sprok** izok jok stok .

6b. wat **dat** krat quat cat .

7a. lalok **farok ororok** lalok **sprok** izok enemok .

7b. wat **jjat bichat** wat **dat** vat eneat .

8a. lalok **brok anok plok nok** .

8b. iat **lat pippat rrat nnat** .

9a. **wiwok nok** izok kantok **ok-yurp** .

9b. **totat nnat** quat oloat **at-yurp** .

10a. lalok mok **nok** yorok **ghirok** clok .

10b. wat **nnat** gat mat bat **hilat** .

11a. lalok **nok** crrrok **hihok** yorok **zanzanok** .

11b. wat **nnat arrat** mat **zanzanat** .

12a. lalok rarok **nok** izok **hihok** mok .

12b. wat **nnat** forat **arrat** vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat
farok = jjat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok **hihok ghirok** .

3b. totat **dat arrat** vat **hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok farok** izok stok .

5b. **totat jjat** quat cat .

6a. **lalok sprok** izok jok stok .

6b. **wat dat** krat quat cat .

7a. **lalok farok ororok lalok sprok** izok enemok .

7b. **wat jjat bichat wat dat** vat eneat .

8a. **lalok brok anok plok nok** .

8b. **iat lat pippat rrat nnat** .

9a. **wiwok nok** izok kantok **ok-yurp** .

9b. **totat nnat** quat oloat **at-yurp** .

10a. **lalok** mok **nok** yorok **ghirok** clok .

10b. **wat nnat** gat mat bat **hilat** .

11a. **lalok nok** crrrok **hihok** yorok **zanzanok** .

11b. **wat nnat arrat** mat **zanzanat** .

12a. **lalok** rarok **nok** izok **hihok** mok .

12b. **wat nnat** forat **arrat** vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat
farok = jjat

lalok = wat / iat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

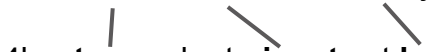
2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok** izok **hihok ghirok** .

3b. totat **dat arrat** vat **hilat** .

4a. **ok-voon anok** drok **brok** jok .



4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok farok** izok stok .

5b. **totat jjat** quat cat .

6a. **lalok sprok** izok jok stok .

6b. **wat dat** krat quat cat .

7a. **lalok farok ororok lalok sprok** izok enemok .

7b. **wat jjat bichat wat dat** vat eneat .

8a. **lalok brok anok plok nok** .

8b. **iat lat pippat rrat nnat** .

9a. **wiwok nok** izok kantok **ok-yurp** .

9b. **totat nnat** quat oloat **at-yurp** .

10a. **lalok** mok **nok** yorok **ghirok** clok .

10b. **wat nnat** gat mat bat **hilat** .

11a. **lalok nok** crrrok **hihok** yorok **zanzanok** .

11b. **wat nnat arrat** mat **zanzanat** .

12a. **lalok** rarok **nok** izok **hihok** mok .

12b. **wat nnat** forat **arrat** vat gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat
farok = jjat

lalok = wat / iat

izok = vat / eneat / quat?
enemok = vat? / eneat?
kantok = quat? / oloat?

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok izok hihok ghirok** .

3b. totat **dat arrat vat hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok farok izok** stok .

5b. **totat jjat quat** cat .

6a. **lalok sprok izok** jok stok .

6b. **wat dat** krat **quat** cat .

7a. **lalok farok ororok lalok sprok izok** enemok .

7b. **wat jjat bichat wat dat** vat eneat .

8a. **lalok brok anok plok nok** .

8b. **iat lat pippat rrat nnat** .

9a. **wiwok nok izok** kantok **ok-yurp** .

9b. **totat nnat** quat oloat **at-yurp** .

10a. **lalok** mok **nok** yorok **ghirok** clok .

10b. **wat nnat** gat mat bat **hilat** .

11a. **lalok nok** crrrok **hihok** yorok **zanzanok** .

11b. **wat nnat arrat** mat **zanzanat** .

12a. **lalok** rarok **nok izok hihok** mok .

12b. **wat nnat** forat **arrat vat** gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat
farok = jjat

lalok = wat / iat

izok = vat / eneat / quat?
enemok = vat? / eneat?
kantok = quat? / oloat?

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok izok hihok ghirok** .

3b. totat **dat arrat vat hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok farok izok** stok .

5b. **totat jjat quat** cat .

6a. **lalok sprok izok** jok stok .

6b. **wat dat** krat **quat** cat .

7a. **lalok farok ororok lalok sprok izok enemok** .

7b. **wat jjat bichat wat dat vat eneat** .

8a. **lalok brok anok plok nok** .

8b. **iat lat pippat rrat nnat** .

9a. **wiwok nok izok** kantok **ok-yurp** .

9b. **totat nnat quat** oloat **at-yurp** .

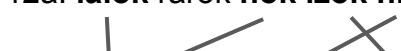
10a. **lalok** mok **nok** yorok **ghirok** clok .

10b. **wat nnat** gat mat bat **hilat** .

11a. **lalok nok** crrrok **hihok** yorok **zanzanok** .

11b. **wat nnat arrat** mat **zanzanat** .

12a. **lalok** rarok **nok izok hihok** mok .



12b. **wat nnat** forat **arrat vat** gat .

13b. iat lat pippat eneat hilat **oloat** at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat
farok = jjat

lalok = wat / iat

izok = vat / quat
enemok = eneat
kantok = oloat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok izok hihok ghirok** .

3b. totat **dat arrat vat hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok farok izok** stok .

5b. **totat jjat quat** cat .

6a. **lalok sprok izok** jok stok .

6b. **wat dat** krat **quat** cat .

7a. **lalok farok ororok lalok sprok izok enemok**

7b. **wat jjat bichat wat dat vat eneat** .

8a. **lalok brok anok plok nok** .

8b. **iat lat pippat rrat nnat** .

9a. **wiwok nok izok kantok ok-yurp** .

9b. **totat nnat quat oloat at-yurp** .

10a. **lalok** mok **nok** yorok **ghirok** clok .

10b. **wat nnat** gat mat bat **hilat** .

11a. **lalok nok** crrrok **hihok** yorok **zanzanok** .

11b. **wat nnat arrat** mat **zanzanat** .

12a. **lalok** rarok **nok izok hihok** mok .

12b. **wat nnat** forat **arrat vat** gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. ???

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat
farok = jjat

lalok = wat / iat

izok = vat / quat
enemok = eneat
kantok = oloat

Centauri-Arcturan Translation

1a. **ok-voon ororok sprok** .

1b. **at-voon bichat dat** .

2a. **ok-drubel ok-voon anok plok sprok** .

2b. **at-drubel at-voon pippat rrat dat** .

3a. erok **sprok izok hihok ghirok** .

3b. totat **dat arrat vat hilat** .

4a. **ok-voon anok** drok **brok** jok .

4b. **at-voon** krat **pippat** sat **lat** .

5a. **wiwok farok izok** stok .

5b. **totat jjat quat** cat .

6a. **lalok sprok izok** jok stok .

6b. **wat dat** krat **quat** cat .

7a. **lalok farok ororok lalok sprok izok enemok**

7b. **wat jjat bichat wat dat vat eneat** .

8a. **lalok brok anok plok nok** .

8b. **iat lat pippat rrat nnat** .

9a. **wiwok nok izok kantok ok-yurp** .

9b. **totat nnat quat oloat at-yurp** .

10a. **lalok** mok **nok** yorok **ghirok** clok .

10b. **wat nnat** gat mat bat **hilat** .

11a. **lalok nok** crrrok **hihok** yorok **zanzanok** .

11b. **wat nnat arrat** mat **zanzanat** .

12a. **lalok** rarok **nok izok hihok** mok .

12b. **wat nnat** forat **arrat vat** gat .

13b. iat lat pippat eneat hilat oloat at-yurp .

13a. lalok brok anok ghirok enemok kantok ok-yurp .

ghirok = hilat

ok-voon = at-voon
ok-drubel = at-drubel
ok-yurp = at-yurp
zanzanok = zanzanat

sprok = dat
ororok = bichat
anok = pippat
plok = rrat
hihok = arrat
nok = nnat
brok = lat
wiwok = totat
farok = jjat

lalok = wat / iat

izok = vat / quat
enemok = eneat
kantok = oloat

It’s actually English-Spanish!

1a. Garcia and associates .

1b. Garcia y asociados.

2a. Carlos Garcia has three associates .

2b. Carlos Garcia tiene tres asociados .

3a. his associates are not strong .

3b. sus asociados no son fuertes .

4a. Garcia has a company also .

4b. Garcia tiene tambien una empresa .

5a. its clients are angry .

5b. sus clientes están enfadados .

6a. the associates are also angry .

6b. los asociados tambien están enfadados .

7a. the clients and associates are enemies .

7b. los clientes y los asociados son enemigos .

8a. the company has three groups .

8b. la empresa tiene tres grupos .

9a. its groups are in Europe .

9b. sus grupos están en Europa .

10a. the modern groups sell strong pharmaceuticals .

10b. los grupos modernos venden medicinas fuertes .

11a. the groups do not sell zanzanine .

11b. los grupos no venden zanzanina

12a. the small groups are not modern .

12b. los grupos pequeños no son modernos

13b. la empresa tiene enemigos fuertes en Europa .

13a. the company has strong enemies in Europe .

Kevin Knight, “Automating Knowledge Acquisition for Machine Translation”
https://www.aaai.org/ojs/index.php/aimagazine/article/download/1323/1224

Two approaches to machine translation

2. Use a lot of translation examples to automatically learn to translate new (unseen) sentences

Two approaches to machine translation

2. Use a lot of translation examples to **automatically learn**??? to translate new (unseen) sentences

Machine Learning

- There are some tasks that
  - We can solve, but
  - We cannot explain how
    - i.e. cannot describe algorithmically
- For example:
  - Spam filtering
  - Optical character recognition (OCR)
  - Translation
  - Face recognition
  - ...

Machine Learning

Instead of explaining how, show how:

- Provide a lot of examples
  - E-mail texts + their classifications (spam/“ham”)
  - Character images + character codes
  - Text in source language + its translation into the target language
- Find patterns and generalize to predict the output of unseen examples

Machine learning

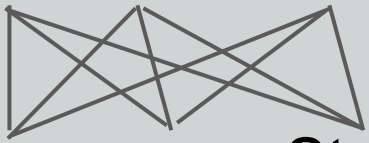
= automatic learning = statistical learning
= statistical modelling = …

- based on some given data
  - like translation examples
- learn to recognize / process new unseen data samples
  - like translate new sentences

- machine learning = generalization
- machine learning = pattern discovery

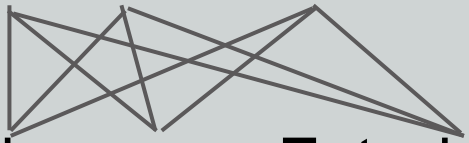
Learning word meanings

the large chair



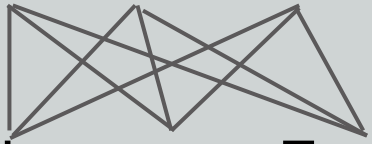
der grosse Stuhl

the big decision



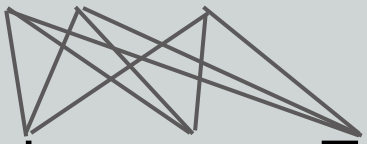
die grosse Entscheidung

the big question



die grosse Frage

a big guy



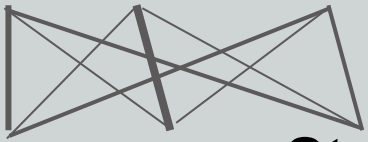
ein grosser Typ

a mistake

ein Fehler

Learning word meanings

the large chair

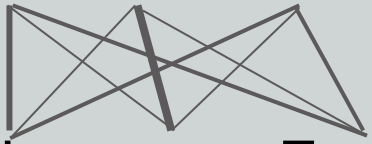


der grosse Stuhl

the big decision

die grosse Entscheidung

the big question



die grosse Frage

a big guy

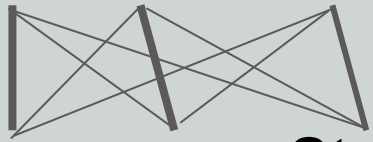
ein grosser Typ

a mistake

ein Fehler

Learning word meanings

the large chair

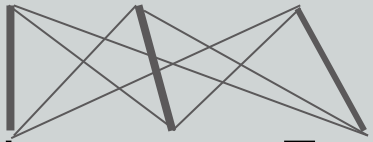


der grosse Stuhl

the big decision

die grosse Entscheidung

the big question

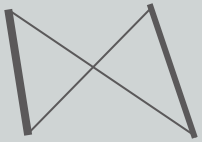


die grosse Frage

a big guy

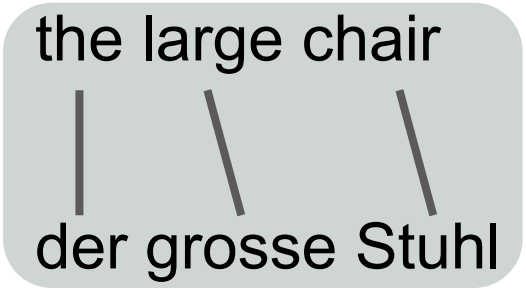
ein grosser Typ

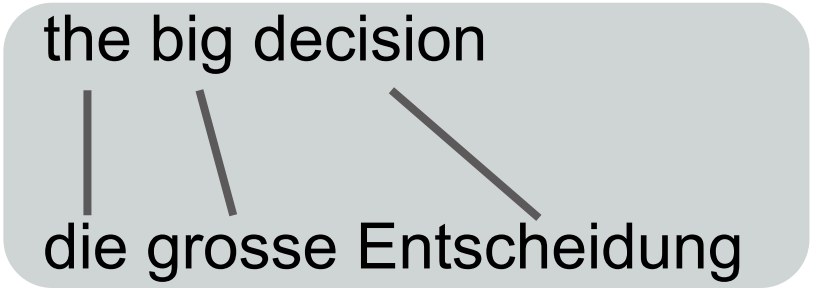
a mistake

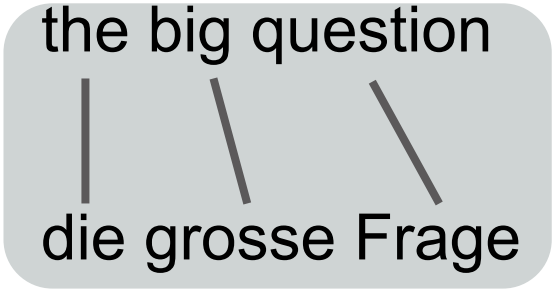


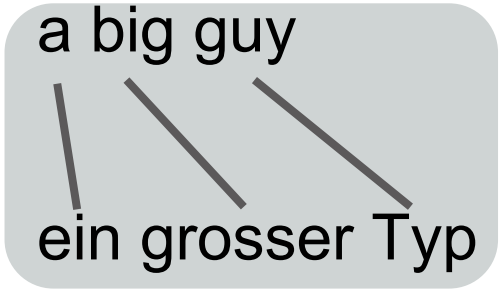
ein Fehler

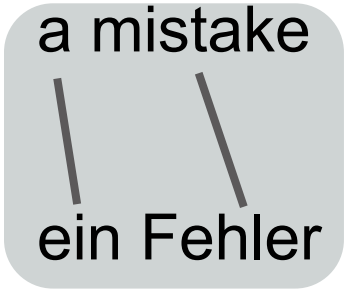
Learning word meanings











Machine translation in practice

- take translation examples for any two language pairs and make a translation system
- any languages
  - modern Estonian → medieval Estonian
  - complicated text → simpler text
- and text type
  - legal texts, technical manuals, subtitles
  - a translation system trained to translate legal texts won't be good at translating news or movie subtitles

Machine translation in practice

- Need texts with translations
  - OpenSubs
  - Europarl
  - TED talks
  - …
- Need a program for learning and translating
  - Moses
    - create translation models (word and phrase alignment and probabilities)
    - create language models
    - translate new texts

Why do machine translation?

- Translations funny and often wrong

Why do machine translation?

- Translations funny and often wrong

Option-1: fix it

Why do machine translation?

- Translations funny and often wrong

Option-1: fix it

- have been trying since the 50ies

Why do machine translation?

- Translations funny and often wrong

Option-1: fix it

- have been trying since the 50ies

Option-2: explain how it's not bug but a feature

- find a way to use is at the quality level that we have

Why do machine translation?

- fixing automatic translations = “post-editing”
- if the translation system is good enough, then automatic translation + fixing is in average faster than manual translation
- take texts of the same kind and let the system learn from them
- post-editing: most widely spread use-case for machine translation

Why do machine translation?

- Google Translate and Bing Translator -- not for post-editing, but for browsing
  - getting the general meaning
- Both use the statistical approach to machine translation
- Google/Microsoft have lots of data = smarter and better system
- Which texts = any texts = harder
  - possible to build a better system for a specific kind of texts, that will be better only for those texts

What is good MT?

- **Correct:** do not buy this product, it is their craziest invention
- **System:** do buy this product, it is their craziest invention

What is good MT?

- **Correct:** do not buy this product, it is their craziest invention
- **System:** do buy this product, it is their craziest invention
  - easy to fix, opposite meaning

What is good MT?

- **Correct:** do not buy this product, it is their craziest invention
- **System:** do buy this product, it is their craziest invention
  - easy to fix, opposite meaning

- **Correct:** The battery lasts 6 hours and it can be fully recharged in 30 minutes.
- **System:** Six hour accumulator, to recharge totally half-an-hour takes

What is good MT?

- **Correct:** do not buy this product, it is their craziest invention
- **System:** do buy this product, it is their craziest invention
  - easy to fix, opposite meaning

- **Correct:** The battery lasts 6 hours and it can be fully recharged in 30 minutes.
- **System:** Six hour accumulator, to recharge totally half-an-hour takes
  - nightmare to fix, meaning: comprehensible

Language technology @ATI

Collaborations:

- translation companies want to try post-editing with machine translation
  - and we can make it for them!
- newspapers want texts analyzed and information extracted
  - we can do that for them too!
- online forums want to automatically detect trolls
  - and we can do that also!

Language technology @ATI

Projects:

- Applications
  - machine translation
  - dialogues
  - summarization
- Components
  - morphological/syntactic/semantic analysis/synthesis
  - corpus processing

Language technology @ATI

Courses:

- Keeletehnoloogia
  - compulsory, 2nd/3rd bachelor year
- Machine translation
- Computational semantics
- Natural language processing in Python
- Syntactic theories and models
  - all non-compulsory subjects


That's all Folks!