Parallele Korpora

Universität München, CIS

“Korpus- und UNIX-Tools”
Sommersemester 2006
Dozenten: Sebastian Nagel, Yeong Su Lee

Natalya Shupletsova, Stefan Partusch

Parallele Korpora

Was ist ein paralleler Korpus?

Sammlung von **identischen** Texten

in **verschiedenen** Sprachen:

- zweisprachig (bilingual)
- mehrsprachig (multilingual)

Was ist ein paralleler Korpus?

Die Idee einen Text parallel in mehreren Sprachen anzubieten ist nicht neu.

A Hec noia filiorū israel: qui in
ταῦτα τὰ ὀνόματα τῶν υἱῶν ἰσραήλ, τῶν εἰσπε
Gen. 46.a. gressi sunt in egyptū cū iacob prē eorū: vnus
πορευμένων εἰς αἴγυπτον ἅμα ἰακὼβ τῷ πατρὶ αὐτῶν, ἕ
quisq; cū domib⁹ suis introierūt: rubē: symeō: leui:
καστος πανοικὶ αὐτῶν εἰσήλθοσαν, ῥουβίμ, συμεών, λευὶ
iudas: isachar: zabulō: ⁊ beniamin: dā: et neph
ἰούδας, ἰσάχαρ, ζαβουλών, καὶ βενιαμίν, δάν, καὶ νεφ
thali: gad: ⁊ aser. ioseph āt erat in egypto. erāt āt
θαλί, γάδ, καὶ ἀσήρ. ἰωσὴφ δὲ ἦν ἐν αἰγύπτῳ. ἦσαν δὲ
oēs aīe q egresse sūt ex iacob: qnq; ⁊ sep
πᾶσαι ψυχαὶ αἱ ἐξελθοῦσαι ἐξ ἰακώβ, πέντε καὶ ἑβ
tuaginta. mortuus ēāt ioseph ⁊ oēs fra
δομήκοντα. ἐτελεύτησε δὲ ἰωσὴφ καὶ πάντες οἱ ἀδελ
tres eius: ⁊ oīs gnātio illa. at filij israel cre
φοὶ αὐτοῦ, καὶ πᾶσα ἡ γενεὰ ἐκείνη. οἱ δὲ υἱοὶ ἰσραὴλ ηὐξή
uerūt: ⁊ multiplicati sūt: ⁊ abūdātes fuerūt: ⁊ in
θησαν, καὶ ἐπληθύνθησαν, καὶ χυδαῖοι ἐγένοντο, καὶ κα
ualuerūt valde nimis. multiplicauit āt tra illos. sur
τίσχυον σφόδρα σφόδρα. ἐπλήθυνε δὲ ἡ γῆ αὐτούς. ἀνέ
rexit āt rex alter sup egyptū: q nō cognoscebat
στη δὲ βασιλεὺς ἕτερος ἐπ' αἴγυπτον, ὃς οὐκ ᾔδει τὸν
ioseph. dixit āt genti sue: ecce gens fi
ἰωσήφ. εἶπε δὲ τῷ ἔθνει αὐτοῦ, ἰδοὺ τὸ ἔθνος τῶν υἱ
liorū israel magna valde multitudo: ⁊ pualet sup nos.
ῶν ἰσραὴλ μέγα πολὺ πλῆθος, καὶ ἰσχύει ὑπὲρ ἡ
venite ergo sapiēter opprimam⁹ eos: ne forte
μᾶς. δεῦτε οὖν κατασοφισώμεθα αὐτούς, μή ποτε
multiplicet: ⁊ qn acciderit nobis bellū: ad
πληθυνθῇ, καὶ ἡνίκα ἂν συμβῇ ἡμῖν πόλεμος, προ
dētur ⁊ isti ad aduersarios. ⁊ de
στεθήσονται καὶ αὐτοὶ πρὸς τοὺς ὑπεναντίους. καὶ ἐκ
bellantes nos: egredietur de tra. ⁊ pre
πολεμήσαντες ἡμᾶς, ἐξελεύσονται ἐκ τῆς γῆς. καὶ ἐπέ
fecit eis pfectos operū: vt affligerēt
στησεν αὐτοῖς ἐπιστάτας τῶν ἔργων, ἵνα κακώσωσιν
B eos in operib⁹. ⁊ edificauerūt ciuitates munitas
αὐτοὺς ἐν τοῖς ἔργοις. καὶ ᾠκοδόμησαν πόλεις ὀχυρὰς
pharaōi: et phithō et ramesses: ⁊ on: q est
τῷ φαραώ, τήν τε πιθὼμ καὶ ῥαμεσσή, καὶ ὤν, ἥ ἐστιν ἡ
heliopolis. quāto āt eos humiliabāt: tāto
λιούπολις. καθ' ὅτι δὲ αὐτοὺς ἐταπείνουν, τοσούτῳ
plures fiebant. ⁊ inualuerūt valde. ⁊ abhominatio
πλείους ἐγίνοντο. καὶ ἴσχυον σφόδρα. καὶ ἐβδελύσσον
nē hūerūt egyptij a filijs israel. ⁊ oppresse
το οἱ αἰγύπτιοι ἀπὸ τῶν υἱῶν ἰσραήλ. καὶ κατεδυνά
rūt egyptij filios israel vi: ⁊ afflixe
στευον οἱ αἰγύπτιοι τοὺς υἱοὺς ἰσραὴλ βίᾳ, καὶ κατωδύ
rūt eorū vitā in operib⁹ duris in
νων αὐτῶν τὴν ζωὴν ἐν τοῖς ἔργοις τοῖς σκληροῖς, ἐν τῷ
luto ⁊ lateritio: ⁊ oib⁹ operib⁹ q in a
πηλῷ καὶ τῇ πλινθείᾳ, καὶ πᾶσι τοῖς ἔργοις τοῖς ἐν τοῖς πε
gris: fm oīa opera: qb⁹ in fuitutē redegerāt eos
δίοις, κατὰ πάντα τὰ ἔργα, ὧν κατεδουλοῦντο αὐτούς

Incipit liber hellesmoth quē nos Exodū dicimus.

Hec sūt nōia filiorū israel: qui ingressi sunt in egyptū cū iacob. Singuli cū domibus suis introierunt. Rubē: Symeon: Leui Iudas: isachar: zabulō & Beniamin. Dā Neptalim. Gad: & Aser. Erāt igitur oēs aīe eorū q egressi sunt de femore iacob septuagīta. Ioseph āt in egypto erat. Quo mortuo & vniuersis fribus eius oīq cognatiōe sua: filii israel creuerūt: & quasi germinātes multiplicati sunt: ac roborati nimis īpleuerūt terram. Surrexit īterea rex nouus sup egyptum: qui ignorabat ioseph. & ait ad populū suū. Ecce ppls filiorū isrl multus & fortior nobis est. Venite sapiēter opprimamus eū: ne forte multiplicetur: & si ingruerit cōtra nos bellū: addatur inimicis nrīs: expugnatisq; nobis egrediatur de terra. Preposuitq; eis magistros opus: vt affligerēt eos oneribus. Edificauerūt q vrbes tabernaculorum pharaōi phiton & ramesses. quātoq; opprimebāt eos: tāto magis multiplicabātur: & crescebant. Oderātq; filios israel egyptii: & affligebant illudentes & inuidentes eis: atq; ad amaritudinem perducebant vitam eorum operibus duris luti & lateris: omniq; famulatu quo in terre operibus

וְאֵלֶּה שְׁמוֹת בְּנֵי יִשְׂרָאֵל הַבָּאִים — בוא
מִצְרָיְמָה אֵת יַעֲקֹב אִישׁ וּבֵיתוֹ
בָּאוּ: רְאוּבֵן שִׁמְעוֹן לֵוִי וִיהוּדָה:
יִשָּׂשכָר זְבוּלֻן וּבִנְיָמִן: דָּן וְנַפְתָּלִי
גָּד וְאָשֵׁר: וַיְהִי כָּל נֶפֶשׁ יֹצְאֵי יֶרֶךְ — היה
יַעֲקֹב שִׁבְעִים נָפֶשׁ וְיוֹסֵף הָיָה
בְמִצְרָיִם: וַיָּמָת יוֹסֵף וְכָל אֶחָיו — מות
וְכֹל הַדּוֹר הַהוּא: וּבְנֵי יִשְׂרָאֵל פָּרוּ — פרה
וַיִּשְׁרְצוּ וַיִּרְבּוּ וַיַּעַצְמוּ בִּמְאֹד מְאֹד — שרץ רבה עצם
וַתִּמָּלֵא הָאָרֶץ אֹתָם: וַיָּקָם מֶלֶךְ — מלא קום
חָדָשׁ עַל מִצְרָיִם אֲשֶׁר לֹא יָדַע
אֶת יוֹסֵף: וַיֹּאמֶר אֶל עַמּוֹ הִנֵּה עַם — אמר
בְּנֵי יִשְׂרָאֵל רַב וְעָצוּם מִמֶּנּוּ: הָבָה — עצם יהב
נִתְחַכְּמָה לוֹ פֶּן יִרְבֶּה וְהָיָה כִּי — חכם
תִקְרֶאנָה מִלְחָמָה וְנוֹסַף גַּם הוּא — קרא לחם יסף
עַל שֹׂנְאֵינוּ וְנִלְחַם בָּנוּ וְעָלָה מִן — לחם
הָאָרֶץ: וַיָּשִׂימוּ עָלָיו שָׂרֵי מִסִּים — שום שור מסס
לְמַעַן עַנֹּתוֹ בְּסִבְלֹתָם וַיִּבֶן עָרֵי — ענה בנה עיר
מִסְכְּנוֹת לְפַרְעֹה אֶת פִּתֹם וְאֶת — סכן
רַעַמְסֵס: וְכַאֲשֶׁר יְעַנּוּ אֹתוֹ כֵּן
יִרְבֶּה וְכֵן יִפְרֹץ וַיָּקֻצוּ מִפְּנֵי — רבה פרץ קוץ
בְּנֵי יִשְׂרָאֵל: וַיַּעֲבִדוּ מִצְרַיִם אֶת — עבד
בְּנֵי יִשְׂרָאֵל בְּפָרֶךְ: וַיְמָרְרוּ אֶת — מרר
חַיֵּיהֶם בַּעֲבֹדָה קָשָׁה בְּחֹמֶר — חיה
וּבִלְבֵנִים וּבְכָל עֲבֹדָה בַּשָּׂדֶה אֵת
כָּל עֲבֹדָתָם אֲשֶׁר עָבְדוּ בָהֶם

Inter p. chal.

Ca. j.
וְאִלֵּין שְׁמָהַת בְּנֵי יִשְׂרָאֵל דְּעָלוּ לְמִצְרַיִם עִם יַעֲקֹב גְּבַר וֶאֱנָשׁ בֵּיתֵהּ
עָלוּ: רְאוּבֵן שִׁמְעוֹן לֵוִי וִיהוּדָה: יִשָּׂשכָר זְבוּלוּן וּבִנְיָמִין: דָּן וְנַפְתָּלִי גָּד וְאָשֵׁר: וַהֲוָה
כָּל נַפְשָׁתָא נְפַקֵי יַרְכָּא דְּיַעֲקֹב שִׁבְעִין נַפְשָׁן עִם יוֹסֵף דַּהֲוָה בְּמִצְרָיִם: וּמִית יוֹסֵף וְכָל

Transla. Chal.

Incipit liber Exodus. Cap. j.

Et hec sunt nomina filiorū israel: qui ingressi sunt in egyptū cum iacob: singuli cum viris domus sue introierūt. Ruben: symeon: leui: et iudas: isachar: zabulon: ⁊ beniamin: dan: ⁊ neptali: gad: ⁊ aser. Erantq; oēs aīe egredientiū de femore

Priūtiua chal.

אזל עלל
מות

Polyglotte Bibeln

Polyglotte (mehrsprachige) Bibeln enthalten die biblischen Texte in Griechisch, Latein, Hebräisch und manchmal anderen Sprachen zum Zwecke der Textkritik.

Motivation heute

In der Computerlinguistik sind parallele Korpora besonders interessant für:

- mehrsprachige Lexikographie
- maschinelle Übersetzung

Parallele Korpora

Parallele Korpora unterscheiden sich durch:

- Anzahl der Sprachen (bi-/multilingual)
- Richtung der Alignierung
  - unidirektional (A -> B)
  - bidirektional (A <-> B)
- Art der Alignierung
  - satzbasiert (sentence alignment)
  - wortbasiert (word alignment)

Alignierung

Ziel der Alignierung ist es die korrespondierenden Sätze oder Wörter der verschiedensprachigen Texte zu finden und zuzuordnen.

Dabei können mehrere Sätze/Wörter auch zu einem Satz/Wort zugeordnet werden.

Parallele Korpora: Satzebene

“Gale & Church”-Algorithmus

“A Program for Aligning Sentences in Bilingual Corpora” (1991)
Wiliam Gale und Kenneth Church
(beide AT&T Bell Laboratories)

Parallele Korpora: Satzebene

“Gale & Church”-Algorithmus

Statistischer Ansatz mit der Grundidee, dass die Länge von korrespondierenden Sätzen korreliert.

Der korrespondierende Satz eines langen/kurzen Satzes in Sprache A ist in Sprache B ebenfalls lang/kurz.

Parallele Korpora: Satzebene

“Gale & Church”-Algorithmus

Die Wahrscheinlichkeit für eine Korrelation ergibt sich aus dem Verhältnis der Zeichenlängen der Sätze und deren Varianz.

Absätze müssen klar markiert sein und sich eindeutig entsprechen. Annahme von “Hard Boundaries” (Absätze) und von “Soft Boundaries” (Sätze).

Parallele Korpora: Wortebene

“Clue Alignment”-Algorithmus

“Combining Clues for Word Alignment” (2003)
Jörg Tiedemann
(Universität Uppsala, Schweden)

Parallele Korpora: Wortebene

“Clue Alignment”-Algorithmus

Ebenfalls ein statistischer Ansatz.
Es wird versucht sog. “Clues” zu finden um die Assoziierung von Wörtern in Quell- und Zielsprache zu bestimmen.

“Clues” können dabei z.B. Frequenz, Wortart, Phrasentypen oder die konkrete Wortform sein.

Parallele Korpora: Wortebene

“Clue Alignment”-Algorithmus

- haben die Wörter viele identische Zeichenfolgen?
- haben sie eine ähnliche relative Frequenz?
- treten sie zusammen auf (Kookurrenz)?
- wie groß ist der Positionsunterschied der Wörter?
- sind die Wortarten “kompatibel”?

Parallele Korpora

Satzalignierte Korpora

European Parlament Proceeding
Parallel Corpus 1996-2003

- ca. 20 Millionen Wörter
- ca. 740.000 Sätze pro Sprache
- 11 Sprachen der EU

http://www.isi.edu/~koehn/europarl

Parallele Korpora

Satzalignierte Korpora

Aligned Hansards of the 36th Parliament of Canada

- ca. 1,3 Millionen Wortpaare
- Englisch - Französisch

http://www.isi.edu/natural-language/download/hansard/

Satzalignierte Korpora

OPUS corpus

- übersetzte Web-Texte
- ca. 500.000 Wörter
- 5 Sprachen (FR, ES, SE, DE, JP)
- unidirektional

http://logos.uio.no/opus/

Parallele Korpora

Satzalignierte Korpora

Slovene-English Parallel Corpus

- ca. 1 Millionen Wörter
- Englisch <-> Slowenisch
- bidirektional

http://nl.ijs.si/elan/

Parallele Korpora

Wortalignierte Korpora

CRATER Multilingual Aligned Annotated Crp.

- drei Sprachen: EN, FR, ES
- 3 x 1 Millionen Token
- morphosyntaktisch annotiert,lemmatisiert
- unidirektional

www.comp.lancs.ac.uk/linguistics/crater/corpus.html

Parallele Korpora: Programme

Manatee/Bonito

Manatee (Server) verarbeitet nur vertikalen Text und Bonito (Client) zeigt Konkordanzen zu Querys an. Kann lediglich von bereits alignierten Sätzen Entsprechungen anzeigen.
Parallele Texte sind bei Manatee technisch verschiedene Korpora!

http://www.textforge.cz/

Parallele Korpora: Programme

Uplug

Uplug ist ein Korpus-Manager und kann konsolen- und web-basiert verwendet werden. Kann parallele Texte automatisch auf Satz- und Wortebene alignieren! Führt Preprocessing durch und bindet externe Tagger ein. Zeigt natürlich auch Konkordanzen zu Querys an.

http://stp.ling.uu.se/cgi-bin/joerg/Uplug