Corpus of Modern Yiddish

Improving the Research Possibilities for Textual Artifacts

Introduction:

What is a corpus?

What is a corpus?

- A collection of texts

  (too weak, too broad a definition)

  many linguists have their own corpora they work with…
- What is a modern corpus?

  (towards a more strict definition)

What is a modern corpus?

A collection of texts that:

- is vast (counts millions to hundreds of millions tokens)
- … and representative (covers various language types and usages and, ideally, periods)
- provides flexible means for efficient linguistic queries (in practice, that means elaborated systems of markup and offline parsing / analysis)

Flexible linguistic queries

- Choosing where you query (subcorpus selection)
- Choosing what you query (semantics, lexicon, syntax, morphosyntax, morphology)
- Accessibility and usability (free online access, user-friendly interface, fast reaction times)

Corpus linguistics: two definitions

- Sense 1. Statistics of vast text collections

offline, use of sophisticated methods of analysis, highly technical interfaces;

mostly applied linguistics

- Sense 2. Corpus-based langs description

online, layman's interfaces, descriptively relevant queries;

language teaching → language description

Corpus linguistics: two definitions

Corpus-based research in Sense 2 has stemmed from the corpus linguistics in Sense 1 relatively lately. As a result, many online corpora (Sense 2) were produced by applied corpus linguists (Sense 2).

Recently, new online corpora started to appear – designed by co-operative teams - applied linguists together with descriptive linguists

Role models

Immediate predecessors of CMY:

- Russian National Corpus
- Eastern Armenian National Corpus

earlier reference points:

- British National Corpus
- Czech National Corpus

…

A tool or ideology?

- not an instrument to provide examples for one's assumptions
- … but an instrument to check them, adjust them and, most importantly, disprove them

Corpus vs. introspection

- linguists differ in how keen their introspection may be – but none has absolute introspective capacities
- language variation can not be covered by one intuition (nor even by a collective intuition of a single linguistic team)
- the corpus provides access to the actual language usage

Corpus vs. introspection

Actual usage accent may vary from

- “Let’s check our intuition against the corpus”

to

- “The corpus is the ultimate truth”

John Sinclair (1982-2003) “Trust the text”

Major shifts of the paradigm

- from prescriptive to descriptive attitudes
- from binary to gradual grammaticality judgements
- from single-system view to synchronic and diachronic variation
- from egalitarian to quantitative evaluation of linguistic forms

(Plungian 2009)

Major shifts of the paradigm

In more general terms:

- from structural to sociolinguistic / diachronic prospective
- de la langue à la parole

General criteria for text selection I

- ideal corpus is a balanced multi-purpose corpus usable for
  - linguistic studies
  - cultural studies
  - language teaching
- represents language in the way it is used in everyday life
  - contains written and spoken language of different genres

General criteria for text selection II

- written language distinguished in printed vs. non-printed texts
- printed texts:
  - fiction of various genres
  - drama
  - memoirs and biographies
  - journalism and literary criticism
  - scientific, popular scientific and teaching texts
  - religious and philosophical texts
  - technical texts
  - business and juridical texts

General criteria for text selection III

- non-printed texts:
  - private letters
  - business letters
  - essays from school or university examinations etc.
  - diaries

General criteria for text selection IV

- spoken texts
  - monologue vs. dialogue
  - spontaneous delivery (e.g. friends talking) vs. primed delivery (e.g. chairman's address to the club members)

General criteria for text selection V
Sociolinguistic variables

- gender
- age
- educational level of the author
- dialect variation
- social contexts the text samples were produced in
  - different orthographies of Yiddish

Yiddish Corpus Linguistics?

- two projects to be mentioned
- "Erstellung eines jiddisch-deutschen Wörterbuchs sowie einer Datenbank jiddischer lexikografischer Hilfsmittel"
  - [compilation of a Yiddish-German dictionary and a database of Yiddish lexicographic additive tools]
  - University of Trier
  - texts used for dictionary compilation conserved in an electronic database
- "Historische Syntax des Jiddischen mit transkribiertem Textkorpus zum älteren Jiddisch (HJS)"
  - [Historical Syntax of Yiddish with a transcribed text corpus of Elder Yiddish]
  - Friedrich-Schiller-Universität Jena
  - texts from the 14th century up to about 1850
  - planned to issue a corpus of the texts used
- **so far no text corpus with morphological annotation and search interface available**

Implications for the Corpus of Modern Yiddish

- texts from 1850 until today
- representation of different dialects and varieties of Yiddish
- different Yiddish orthographies (e.g. Soviet Yiddish)
- different text genres

- non-sociolinguistic factors to be considered
- general availability of texts
  - how much text material is available for a given period?
  - is the material well-preserved enough for text processing (e.g. scanning, OCR)?
  - are all text genres available for a given period?
  - balanced corpus possible for a given period?

- original vs. translation
  - translations often influenced in syntax and lexicon by original language
  - translations included only if overall text production in the given genre is small (e.g. texts about natural sciences)

Corpus architecture
Allocation of word forms to time periods

- CMY will consist of two major subcorpora:
  - basic corpus (all text genres) containing 10 million wordforms
  - corpus of modern newspaper texts (as big as possible)
- allocation of wordforms to time periods in the basic corpus
- 2 million wordforms for period 1850-1900
- 6 million wordforms for 1900-1939
- 2 million wordforms for 1940 – today

- period 1900-1939 covers the height of Yiddish
- all text genres and all dialects / varieties available in this period
- language usage of this period serves as model for language teaching
- preservation of texts best for 1900-1939 period
- possibility of compiling a balanced subcorpus 1900-1939

Allocation of word forms 1900-1939
Text genres

- fiction texts, including poetry 3 million wordforms
- newspapers 1.5 million wordforms
- scientific texts, schoolbooks 1 million wordforms
- *Gebrauchstexte*
  (instruction manuals, promotion etc.)
- handbooks, official domain 0.5 million wordforms

- (memoirs, personal narratives and letters, journals and diaries) closest to colloquial style
- difficult to obtain and to process
  - usually in private hands > how much material does exist at all?
  - usually handwritten > difficult to scan and OCR
  - published personal writing of VIPs > adapted to literary standards
- genre “scientific texts, schoolbooks“ may contain translations
  - few original Yiddish texts especially in natural sciences
  - impact of source language common phenomenon for this genre

Allocation of word forms 1900-1939 Geographical factor

- representation of the different geographic (and thus dialectal) varieties of Yiddish
- consideration of the cultural centres of the given time:

North America, including Canada	1.5 mln WF
South America	0.75 mln WF
Soviet Union	0,75 mln WF
Poland and rest of Eastern Europe	2.25 mln WF
Western Europe and rest of the world	0.75 mln WF

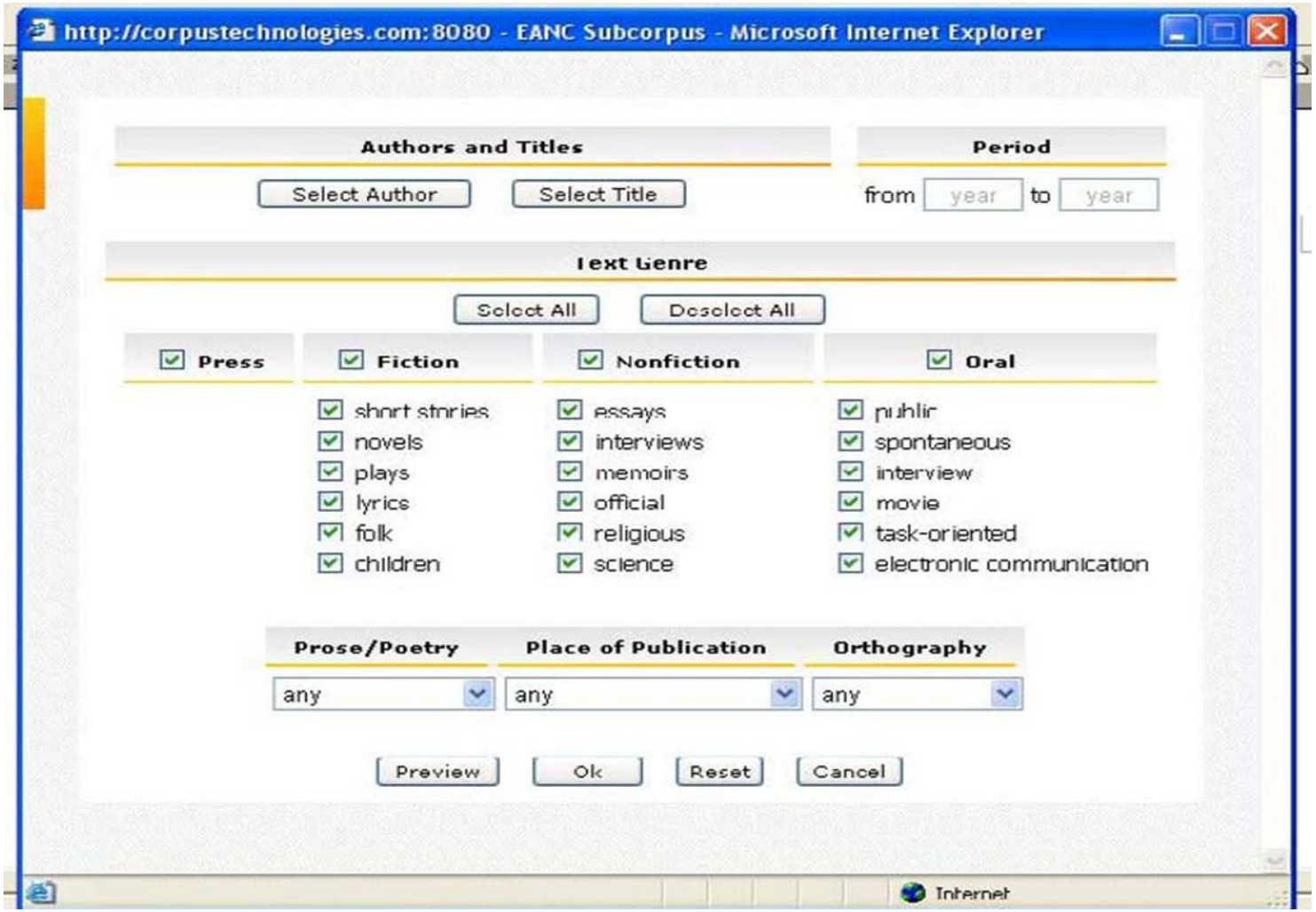
Problems with the geographical factor

- for books place of publication is not a reliable criterion for ascribing dialect status
  - representation of geographic varieties is to be achieved mainly by newspaper texts
- search in dialect subcorpora requires some labels for these subcorpora
  - denominations of countries as subcorpora labels imply historical developments
  - e.g. status of Vilnius?
  - *Language and Culture Atlas of Ashkenazic Jewry* as basis for

    drawing boundaries
  - CMY manual will provide survey of geographical denominations and places ascribed to them

Text annotation
Selection tags

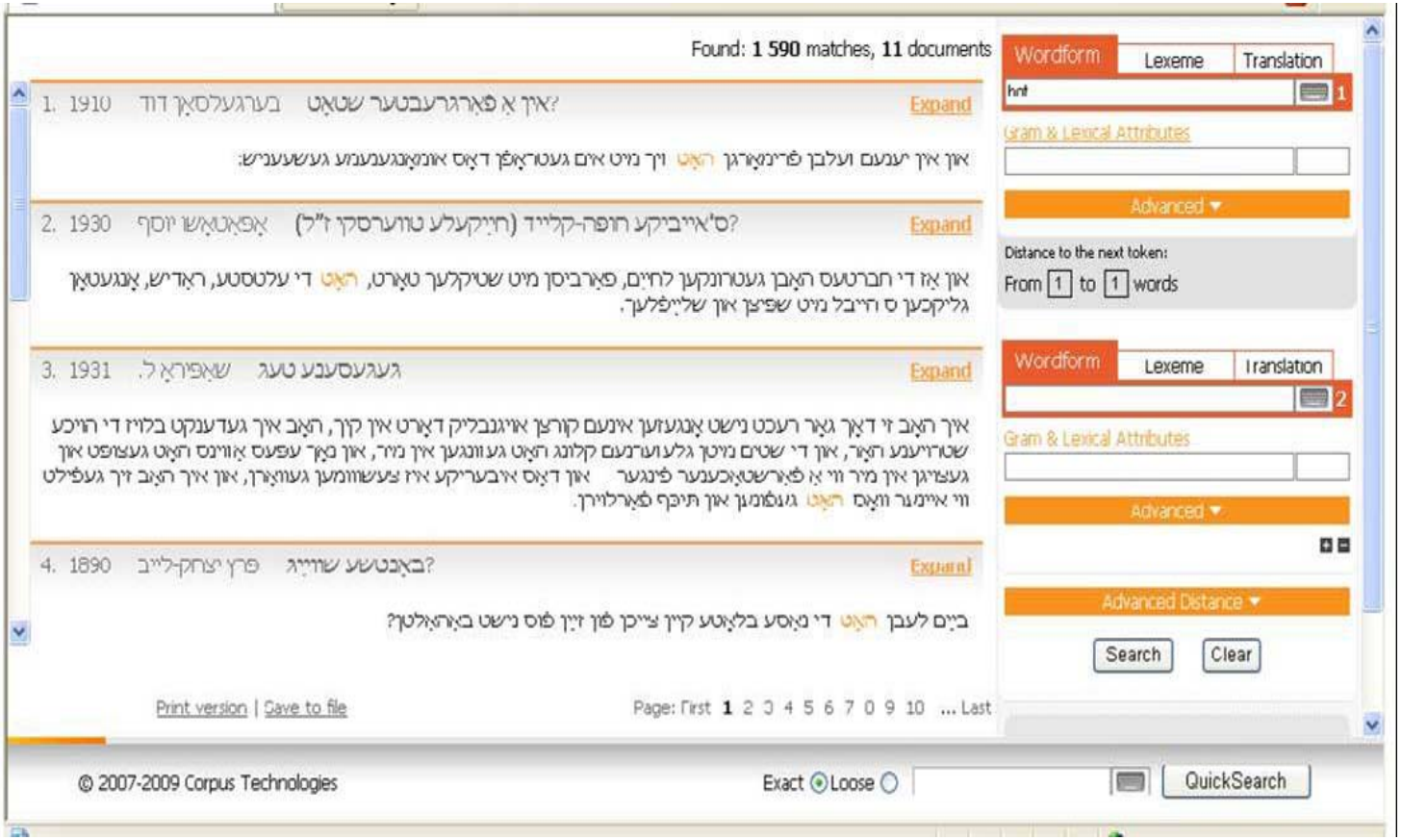
- each text in CMY bears annotation, i.e. tags with various information
- distinction of selection tags and display tags
- selection tags used for subcorpus selection
  - name of the author in Hebrew letters (conventional transliteration of the name into Latin)
  - title (transliteration of the title)
  - year (=year(s) o creation > year of the first pubclication > proxy based on the author's lifespan)
  - genre (e.g. press, fiction, non-fictional texts)
  - subgenre (e.g. fiction > short stories, novels, plays, lyrics)
  - place of publication
  - region of publication
- user may compile personal subcorpus
- e.g. “search all plays written between 1905 and 1920 in Poland“


http://corpustechnologies.com:8080 - EANC Subcorpus - Microsoft Internet Explorer
Authors and Titles
Select Author
Select Title
Period
from year to year
Text Genre
Select All
Deselect All
Press
Fiction
Nonfiction
Oral
short stories
novels
plays
lyrics
folk
children
essays
interviews
memoirs
official
religious
science
public
spontaneous
interview
movie
task-oriented
electronic communication
Prose/Poetry
Place of Publication
Orthography
any
any
any
Preview
Ok
Reset
Cancel
Internet

Text annotation
Display tags

- display tags are attached to the sentences matching the user‘s query
- offer information about the texts from which matches are taken

- name of the author in Hebrew letters (conventional transliteration of the name into Latin)
- title (transliteration of the title)
- year(s) o creation > year of the first publication
- place of publication


Found: 1 590 matches, 11 documents
1. 1910 בערגעלסאָן דוד איך אַ פּאַרגרעבטער שטאָט? Expand
און אין יענעם ועלבן פּ־נאָמען האָט ויך נויט אים געטראָפן דאָס אונטאָגענעמע געשעעניש:
2. 1930 אָפּאָטאָשו יוסף ס'אייביקע חופה-קלייד (חיקעלע טוערסקי ז"ל)? Expand
און אַז די חברטעס האָבן געטרונקען לחיים, פּאַרביסן מיט שטיקלעך טאָרט, האָט די עלטסטע, ראָדיש, אָנגעטאָן גליקכען ס וויבל נויט שפּיצן אין שליפּלעך.
3. 1931 שאָפּיראָ ל. געגעסענע טעג Expand
איך האָב זי דאָך גאָר רעכט נישט אָנגעזען אינעם קורצן אויגנבליק דאָרט אין קך, האָב איך געדענקט בלויז די הויכע שטערעענע האָר, אין די שטים נויטן גלעזערנעם קלונג האָט גע וונגען איך נויט, אין נאָך עפּעס אַווינס האָט געצופט אין געצויגן אין מיר וויי אַ פּאַרשטאָכענער פּינגער און דאָס איבעריקע איז צעשוומען געוואָרן, אין איך האָב זיך געפילט ווי איינער וואָס האָט גענומען אין תיכף פּאַרלוירן.
4. 1890 פרץ יצחק-לייב בראָטשע שוויץ? Expand
ביים לעבן האָט די נאָסע בלאָטע קיין צייכן פון זיין פוס נישט באָהאַלטן?
Print version | Save to file
Page: First 1 2 3 4 5 6 7 8 9 10 ... Last
Wordform Lexeme Translation
hnt 1
Gram & Lexical Attributes
Advanced
Distance to the next token:
From 1 to 1 words
Wordform Lexeme Translation
2
Gram & Lexical Attributes
Advanced
Advanced Distance
Search Clear
© 2007-2009 Corpus Technologies
Exact Loose QuickSearch

What can CMY be used for?

- language teaching
  - instructors may use CMY as resource for natural-language examples (e.g. double negation)
  - students may deduce grammatical rules from CMY samples

    (e.g. conjugation of periphrastic verbs)
- research issues
  - usage-based description of language
  - instrument for checking one‘s hypotheses

Thanks go to

- **Fritz-Thyssen-Foundation**
  for financial support (grant number 10.09.1.065)
- **Binyumen Shekhter Foundation for the Advancement of Yiddish**
  for financial support
- **National Yiddish Book Center**
  for allowing to incorporate texts from their online library
- **Rachel Heuberger** from the Judaica Collection of the University Library Frankfurt / Main
  for allowing to incorporate texts from their online library
- **Raphael Finkel**
  for help with OCRing texts
- **Forverts**
  for allowing to incorporate texts from their website
- **Lebns-fragn**
  for allowing to incorporate texts from their website
- **Hamaspik Gazette**
  for allowing to incorporate texts from their website
- **Afn shvel**
  for allowing to incorporate texts from their website
- **… join the club!**

Team members

Nikita Bezrukov, Sandra Birzer, Michael Daniel, Björn Hansen, Mikhail Kudinov, Holger Nath, Elena Luchina, Alexei Poljakov, Vladimir Plungian, Alexandra Poljan, Evita Wiecki