Contrastive Collocation Analysis – a Comparison of Association Measures across Three Different Languages Using Dependency-Parsed Corpora

Stefan Evert[1], Thomas Proisl[1],
Peter Uhrig[1], Maria Khokhlova[2]

[1] University of Erlangen-Nuremberg

[2] St. Petersburg State University

Approaches to Collocation Extraction

Collocation:

combination of two lexical items as listed in collocations or explanatory dictionaries.

Evert 2004:

- segment-based co-occurrences;
- distance-based co-occurrences;
- **relational co-occurrences**.

Analysis

- English, German, Russian;
- dependency-annotated corpora;
- binary collocations (adj-noun);
- association measures

Data

- DECOW16A (Schäfer & Bildhauer, 2012);
- ENCOW16A (Schäfer, 2015);
- Araneum Russicum II Maximum (Benko, Zakharov, 2016);
- dependency parsed corpora:
  - English: spaCy;
  - German: mate-tools
  - Russian: UDPipe 1.20

Gold Standards

- English: Oxford Collocations Dictionary for Students of English, 2nd ed. 2009 (OCD2);
- German: Wörterbuch der Kollokationen im Deutschen, 2011 (WdK)
- Russian: Dictionary of Russian in 4 vol. (Jevgen'jeva 1981-1984).

Collocations from Gold Standards

- 86,563 adj-noun pairs extracted from the English gold standard;
- Examples:
  - heavy smoker;
  - Western democracy;
  - inner demon;
- 105,872 adj-noun pairs extracted from the German gold standard;
- Examples
  - farbig + Abbildung
  - radioaktiv + Abfall
  - stattlich + Abfindung

Collocations from Gold Standards

- 3,387 adj-noun pairs extracted from the English gold standard;
- Examples:
  - академический год
  - большой палец
  - декретный отпуск
  - медовый месяц
  - периодическая система
  - сухой закон

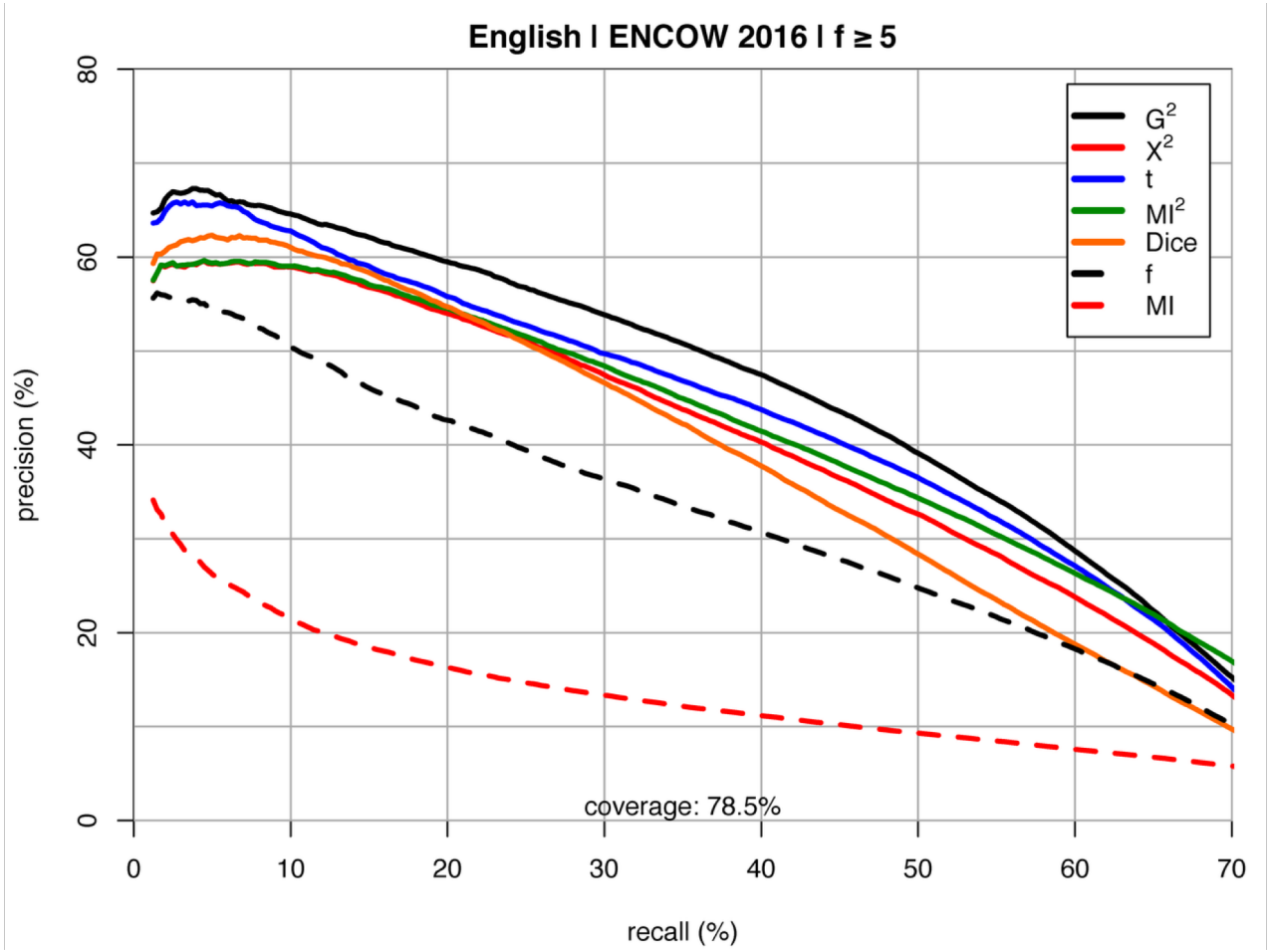
Association Measures

- Main cue: co-occurrence frequency, quantified by various statistical association measures.
- 84 measures (Pecina 2005); 47 measures (Wiechmann 2008).
- Evaluated by (Evert et al., 2017) on a smaller English gold standard:
  - MI;
  - MI^2;
  - t-score;
  - Dice;
  - X^2;
  - log-likelihood

Evaluation

- precision-recall curves;
- precision (P): the percentage of true positives among the n candidates;
- recall (R): the percentage of all true positives in the gold standard found in the n-best list;
- the “higher” a P/R graph is located in the plot, the better the ranking achieved by the corresponding association measure;
- AP50: a composite measure for average precision, recall = 50%

Results for English Adj-Noun Collocations

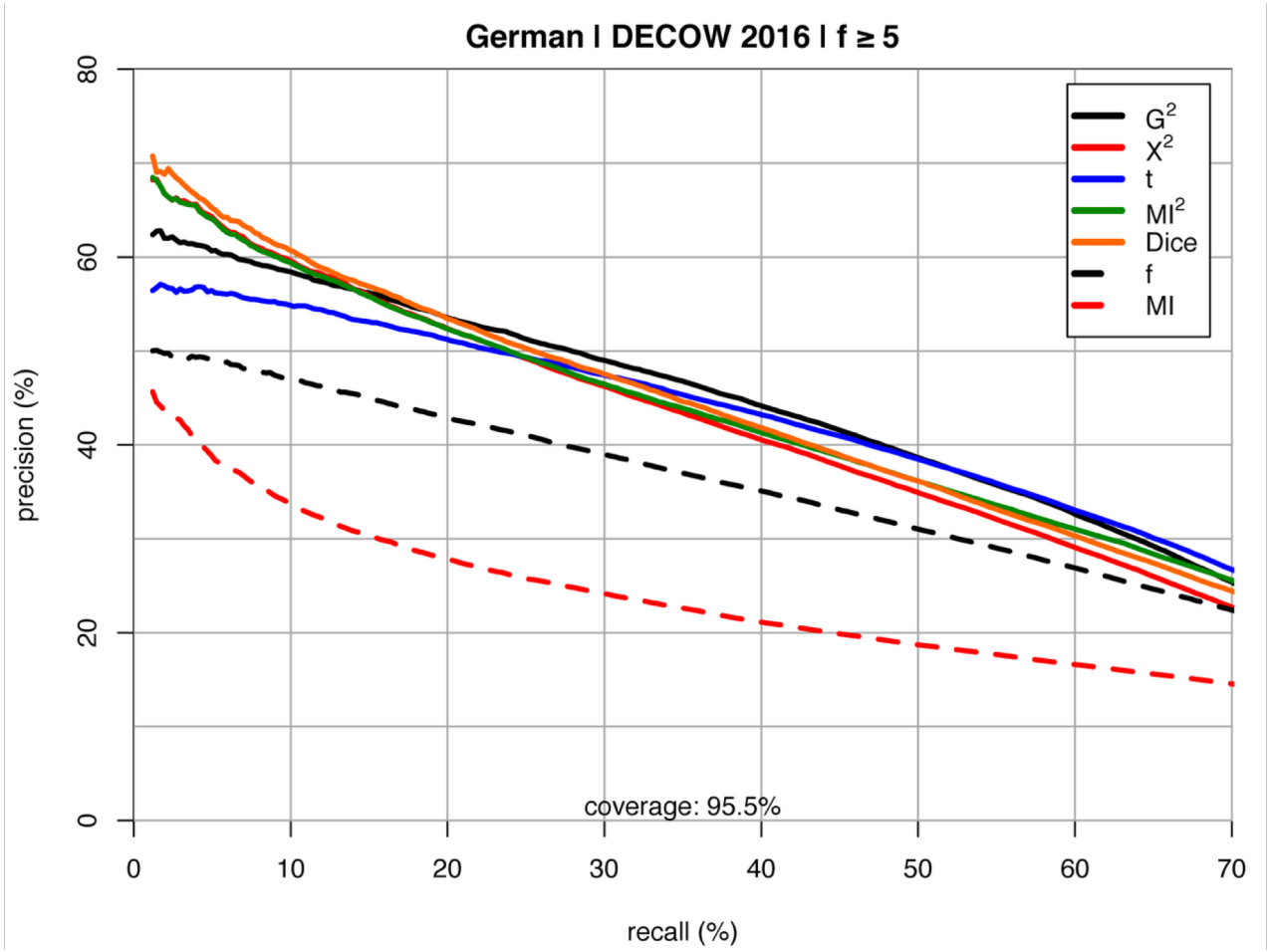


Results for English Adj-Noun Collocations

- the best overall measure: log-likelihood;
- coverage 78.5%

l1	l2	b.TP	G^2
year_N	last_J	TRUE	7115806,250
device_N	electronic_J	TRUE	4454791,250
range_N	wide_J	TRUE	3604840,760
information_N	more_J	FALSE	3527524,720
week_N	last_J	TRUE	3012003,080
party_N	third_J	FALSE	2734559,170
navigation_N	usual_J	FALSE	2657799,180

Results for German Adj-Noun Collocations

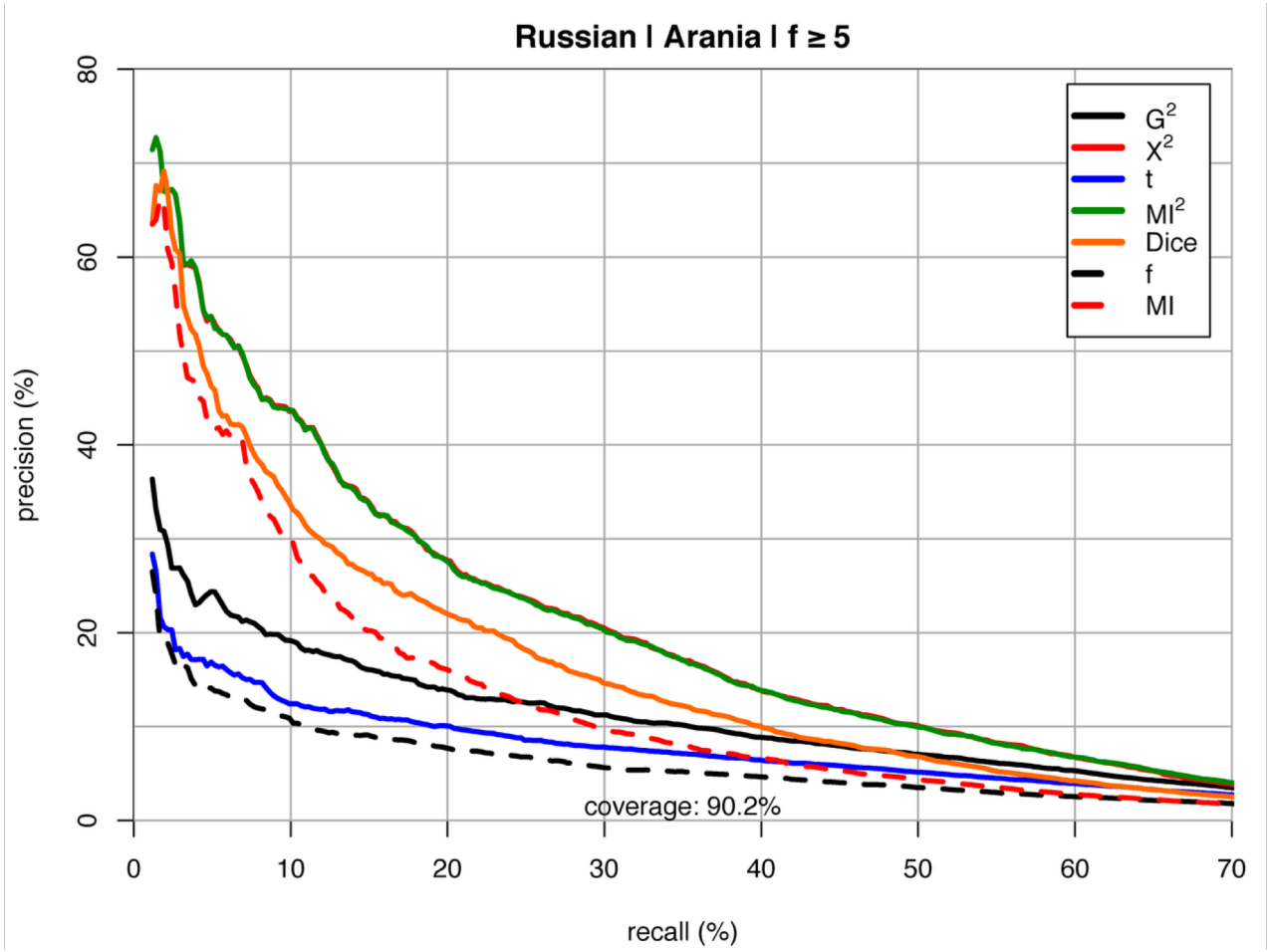


Results for German Adj-Noun Collocations

- the best overall measure: Dice / log-likelihood;
- coverage 95.5%

l1	l2	b.TP	G^2
Jahr_N	vergangen_J	TRUE	3524676,622
Zeit_N	kurz_J	TRUE	2728721,674
Zeit_N	lang_J	TRUE	2341170,448
Jahr_N	nah_J	FALSE	2029583,119
Stunde_N	halb_J	FALSE	1819704,642
Energie_N	erneuerbar_J	TRUE	1685377,270
Tag_N	nah_J	FALSE	1555725,796

Results for Russian Adj-Noun Collocations

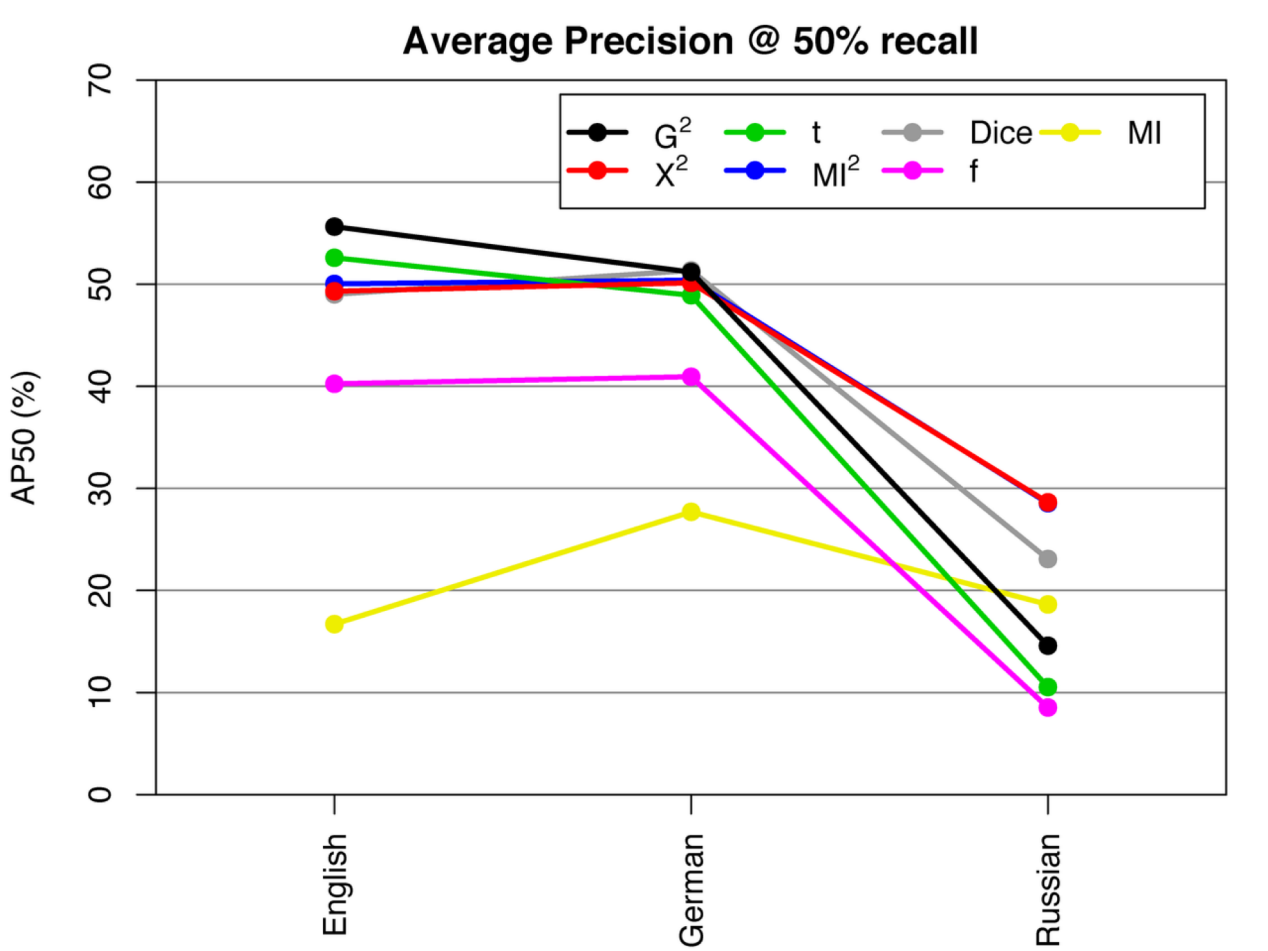


Results for Russian Adj-Noun Collocations

- the best overall measure: MI^2 / X^2;
- coverage 90.2%

l1	l2	b.TP	G^2
дело_N	самый_J	TRUE	10097423,630
день_N	сегодняшний_J	TRUE	8772634,170
директор_N	генеральный_J	FALSE	7090687,977
год_N	прошлый_J	FALSE	6815974,206
участок_N	земельный_J	FALSE	6805317,506
плата_N	заработный_J	TRUE	5713262,133
сад_N	детский_J	TRUE	5176036,660

Average Precision at 50% (AP50)



Conclusions and Further Work

- Log-likelihood is the best association measure for this type of collocation for English and German but not for Russian.
- For some specific types other measures can perform even better.
- Carefully sampled and balanced corpora seem to have considerable advantages in precision.
- Larger, less balanced corpora (such as the web corpora) can be more useful for total coverage.
- Further enlargement of Russian gold standard.

Thank you!