Statystyki korpusu “rapid”

- Korpus składa się z 17000 książek pobranych z rapidshare. Książki są w formacie tekstowym, kodowane głównie w cp1250.
- W korpusie znajduje się ponad 6GB danych, baza n-gramów zajmuje 42GB

Statystyki korpusu “rapid”

- Ilość słów: 1 075 601 165

- Ilość 1-gramów: 7 032 022
- Ilość 2-garmów: 159 614 194
  - 22.7 x więcej niż 1-gramów
- Ilość 3-gramów: 439 272 946
  - 62.5 x więcej 1-gramów

Statystyki korpusu “rapid”

- Ilość uszkodzonych 1-gramów: 601 420 (8%)
- Ilość uszkodzonych słów: 14 069 793 (1.3%)

- 1-gramy występujące 1 raz: 4 119 667 (58%)
- 2-gramy występujące 1 raz: 114 073 963 (71%)
- 3-gramy występujące 1 raz: 368 893 176 (84%)

Statystyki korpusu “rapid”

- 1-gramy występujące więcej niż 10 razy: 951 844
- 2-gramy występujące więcej niż 10 razy: 6 426 198
- 3-gramy występujące więcej niż 10 razy: 5 166 537

- 1-gramy występujące więcej niż 20 razy: 650 497
- więcej niż 50 razy: 415 200
- więcej niż 100 razy: 284 030
- więcej niż 500 razy: 105 032
- więcej niż 1000 razy: 63 940

Statystyki korpusu “rapid”

- Średnia długość 1-gramu: 9.41
- Średnia ważona: 5.14

  za wagę przyjęto ilość wystąpień 1-gramu

- Długość słowa jest ograniczona do 64 znaków

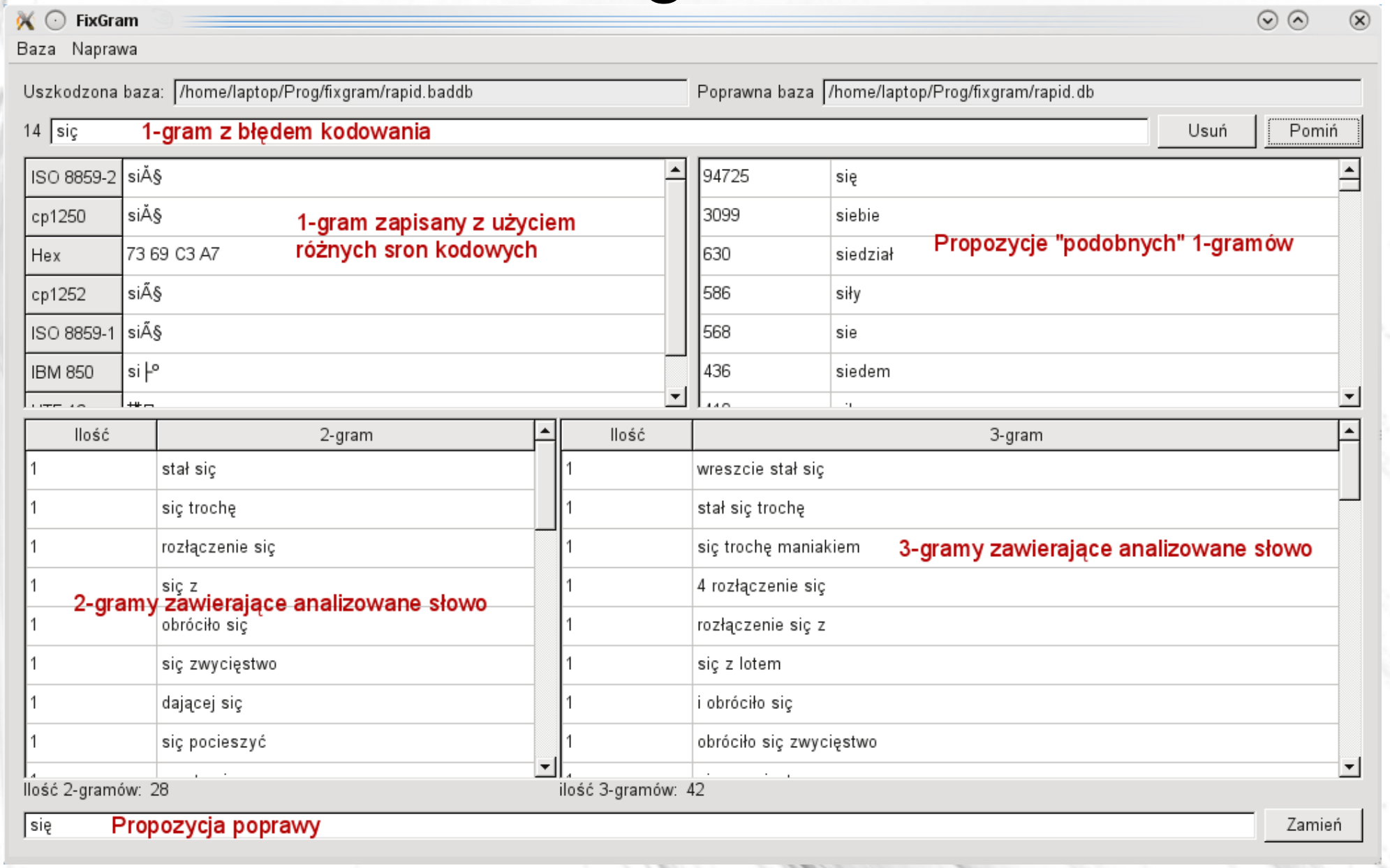
Fixgram

- Program do tworzenia bazy ngramów, jako wynik daje obecnie 2 pliki. Bazę z poprawnymi ngramami oraz bazę zawierającą wyłącznie n-gramy z błędami kodowania.
- Program służy do naprawy błędów kodowania oraz błędów ortograficznych występujących w bazie.
- Podczas poprawy błędów kodowania, możliwa jest praca z bazą zawierająca tylko n-gramy z błędem, lub z obiema bazami.

Fixgram

- Podczas pracy z obiema bazami, korekta jest na bieżąco zapisywana w poprawnej bazie. W wypadku pracy tylko z bazą zawierającą błędy kodowania, korekta może być włączona do poprawnej bazy w dowolnym momencie.
- Każdy proces naprawy bazy może być zapisany w pliku i odtworzony dla innej bazy. Dzięki temu człowiek nie musi 2 razy naprawiać tego samego błędu.

Fixgram


FixGram
Baza Naprawa
Uszkodzona baza: /home/laptop/Prog/fixgram/rapid.baddb
Poprawna baza /home/laptop/Prog/fixgram/rapid.db
14 siç
1-gram z błędem kodowania
Usuń
Pomiń
ISO 8859-2 siÄ§
cp1250 siÄ§
Hex 73 69 C3 A7
cp1252 siÃ§
ISO 8859-1 siÃ§
IBM 850 si├º
1-gram zapisany z użyciem różnych sron kodowych
94725 się
3099 siebie
630 siedział
586 siły
568 sie
436 siedem
Propozycje "podobnych" 1-gramów
Ilość 2-gram
1 stał siç
1 siç trochę
1 rozłączenie siç
1 siç z
1 obróciło siç
1 siç zwycięstwo
1 dającej siç
1 siç pocieszyć
2-gramy zawierające analizowane słowo
Ilość 3-gram
1 wreszcie stał siç
1 stał siç trochę
1 siç trochę maniakiem
1 4 rozłączenie siç
1 rozłączenie siç z
1 siç z lotem
1 i obróciło siç
1 obróciło siç zwycięstwo
3-gramy zawierające analizowane słowo
Ilość 2-gramów: 28
ilość 3-gramów: 42
się
Propozycja poprawy
Zamień