RO-Stemmer mit Snowball

Spezifikation

17.05.2006

Doina Gligă
Erwin Glockner
Marina Stegărescu

Inhaltsübersicht

- Stemmer
  - Porters Algorithmus
  - Porters Snowball
- Rumänisch
  - Flexionsstruktur
  - Homonymie
- Implementierung
  - Werkzeug
  - Ablauf

Was wollen wir machen?

- Entwicklung eines Stemmers in Snowball für Rumänisch

Was sind Stemmer?

- Programme, die Wörter auf ihren gemeinsamen Wortstamm zurückführen
- vor allem als Komponente der IR-Systeme entwickelt und benutzt
- Linguistische Analyse

Stemmer

- Lexikonbasierte
- Korpusbasierte
- Regelbasierte

Was braucht man für die Entwicklung eines Stemmers?

- Verfahren → Porters Algorithmus
- Sprache → Snowball

Porters Algorithmus

- Menge von Verkürzungsregeln: Bedingungen und Ableitungen für verschiedene Suffixe
- Maßgeblich: Vokal-Konsonant-Sequenzen
- Regelanwendung auf die Vokal-Konsonant-Sequenzen

Snowball

- Snowball: stringverarbeitende Sprache
- ermöglicht das einfache und exakte Repräsentieren von Stemmingalgorithmen
- entwickelt von Martin Porter

Porters Idee

- Entwicklung einer Sprache " … in which the rules of stemming algorithms can be expressed in a natural way."

Snowball vs. C

loop AE C

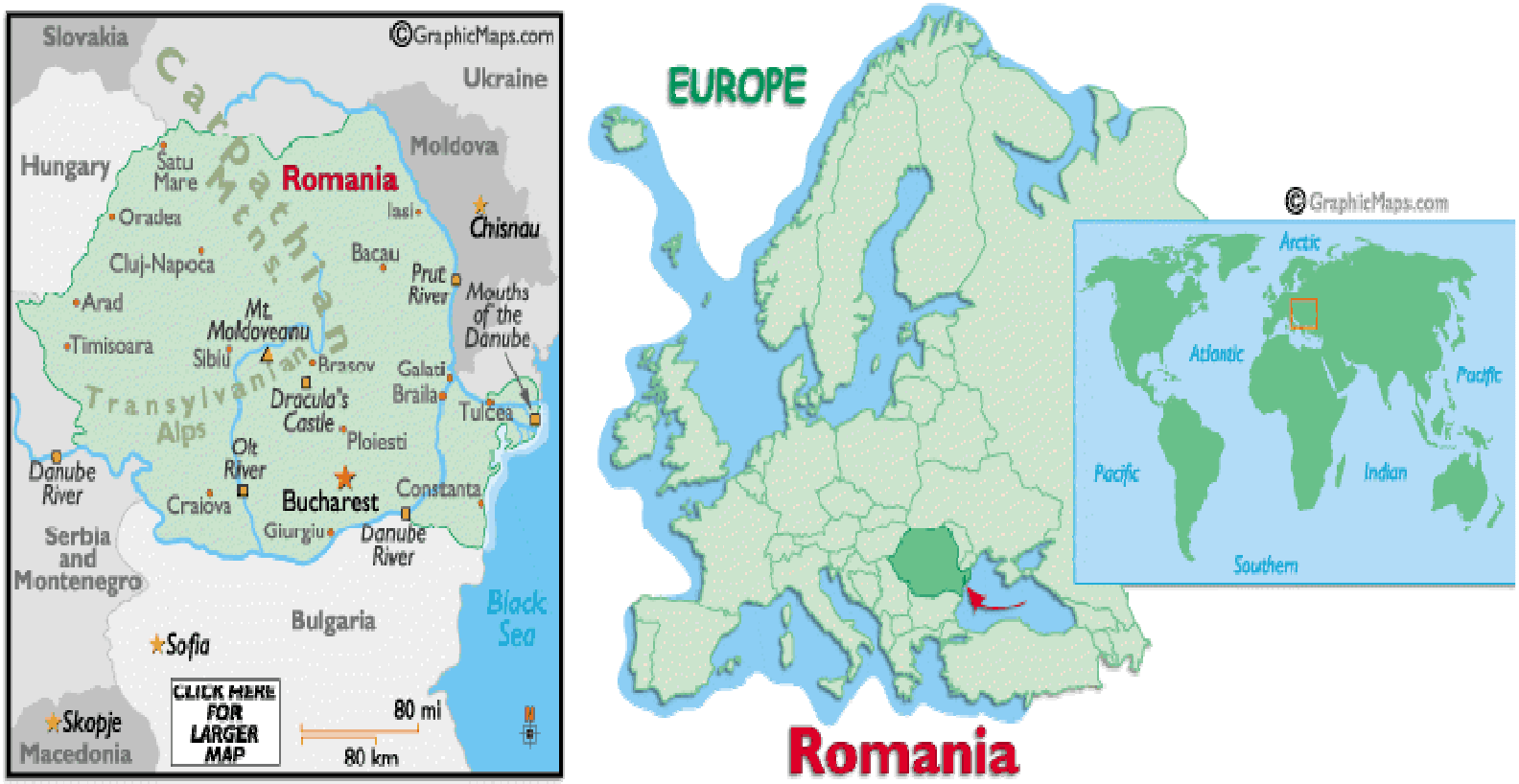
This is like C C ... C written out AE times, where AE is an arithmetic expression. For example,

```
$x loop 2 gopast ('a' or 'e' or 'i' or 'o' or 'u')
   /* position c after the second vowel */
```

The equivalent expression in C has the shape,

```
{    int i;
        int limit = AE;
        for (i = 0; i < limit; i++) C;
  }
```

Rumänien



http://worldatlas.com/webimage/countrys/europe/ro.htm

Das Rumänische

- Romanische Sprache, mit einem starken slavischen Einfluß
- Wortschatz:
  - Ca. 70% romanischer Herkunft (Lateinisch + andere romanischen Sprachen)
  - Ca. 20% - Slavisch
  - ~10% anderer Herkunft (Türkisch, Ungarisch, Griechisch, Deutsch etc)
- Das grammatische System - > lateinisch
- 7 Vokalen <a,e,i,ă,â/î,o,u>
- 22 Konsonnanten <ş, ţ>

Begriffe

- Wurzel (Root) - die Sequenz des Wortes, die nicht mehr zerlegbar ist und in lautlicher und semantischer Hinsicht, als Ausgangsbasis entsprechender Wortfamilie angesehen wird
  - Cânt-a (singen)
- Stamm – Morphem oder Morphemkonstruktion, an die Flexionsendungen treten können
  - Descânt-a (durch Spüche Zauber vertreiben)
- Flexionselemente – die Menge aller Elemente, die in paradigmatischer Relation mit dem Stamm eines Wortes sind, und das

  Flexionsparadigma dieses Wortes bildet
- Flexionsparadigma – die Menge aller Flexionsformen des Wortes

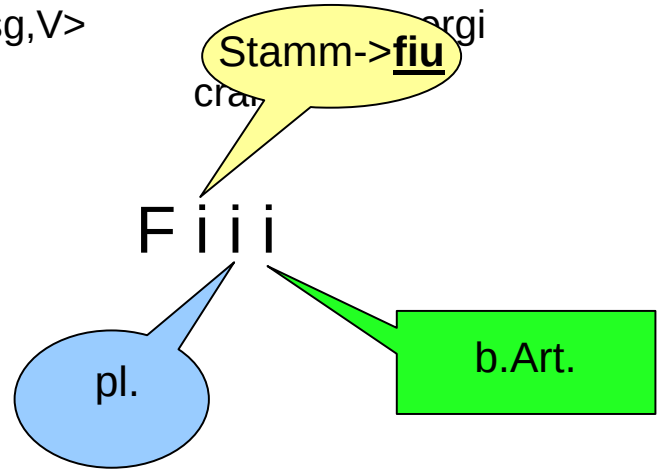
Rumänische Morphologie

- Das Wort – 1 – 7 Silben; (Flektionsmarker inkl.)
  - Pori [1] (Poren)
  - Imbunatatirile [6] (die Verbesserungen)
- 1-3 Stämme
  - Pom <sg, o.Art> pom -i <pl., o.Art> pom-u-lui <sg. Art. G/D>
  - Fat-a fet-e
  - Om <sg, o.Art> oamen -i
  - Frumos <sg., m.> frumoş -i <pl.m.> frumoas-e <pl. f.>
- Flexionsstruktur - umfangreich
  - **Nom** : - Genus: <M,F,N>>
    - Art <+bestimmter: -> (+/-Des) Suffix> <unbestimmter :-> anderes Wort>
    - Numerus <Sg, Pl>
    - Casus <N,Ak,D,G,V>
  - **Adjektiv** : <+ Art : Adj + N>
  - **Verb**: <Gruppe: 4>, Modus: <Präd: 4><Npräd:4>
- P:: Stamm +(Vok)+ (Suffix) + (Suffix) + Flektionsmarker
  - Muncitorimea

Probleme

- Homonymie der Endungen:
  - -e: <+pl, N> case (Häuser)
    <+inf, V> merge (gehen)
    <+stamm> bine (gut)
  - -i <+pl,N,Adj> copaci (bäumer) frumoşi (schöne (Kinder)
    <+b.Art,N> iubi (lieben)
    <+inf, V> [illegible]rgi (du gehest)
    <+2.Pers,sg,V> cra[illegible] (Prinz)
    <+stamm>

-ar <Suffix> familiar
< +stamm> marar (Dill)



Literatur

- Luciana Peev, Lidia Bibolar, Jodal, Endre, **A Formalization Model of the Romanian Morphology**
- *http://www.racai.ro/books/awde/peev.html*
- Jörg Meibauer & al. , **Einführung in die germanistische Linguistik**, Stuttgart, 2002
- I. Coteanu, **Limba română contemporană**, vol. I, , Bucureşti, 1974
- *http://snowball.tartarus.org/*