Project Python & Woordenschatverwerving

**Algemene doelstelling**

Toepassingen in het kader van woordenschatverwerving.

De deelprojecten kunnen telkens werken met een (zeer) beperkte selectie van lexicale elementen: het gaat om de codeerprincipes, niet om de hoeveelheid data die al verwerkt worden

**Deelproject 1: Nieuwe tekstbronnen scannen en analyseren in lemma’s en/of n-grams** (met behulp van lemmatiseerder TreeTagger)

1. input
   1. mogelijkheden
      1. website aanduiden en automatisch downloaden in propere format
      2. tekst copy-pasten in WORD
      3. tekst opslaan in utf-8
   2. pitfalls
      1. brontekst niet in juiste format (automatisch converteren naar utf-8 of foutmelding geven)
      2. codeerresten vanuit copy-paste uit html bestanden
      3. anticiperen op tokenisatie problemen
2. proces
   1. TreeTagger automatisch aanroepen vanuit prompt
3. output
   1. format om weg te schrijven in leesbare tekst met tags
4. testmogelijkheden
   1. teksten uit convergentiecorpus

**Deelproject 2: Classificeren van lemma’s volgens categorieën van geziene / nieuwe woordenschat of woordenschat niveau A1-A2-B1-B2-C1**

1. Technische filtering van lemma’s uit TreeTagger output
   1. alleen uit lexicale woordklassen (geen symbolen, interpunctie, functiewoorden)
   2. geen <unknown>
2. Didactische filtering volgens verschillende referentiepunten
   1. eigen referentiepunt UGent
      1. woorden uit reeds geziene basiswoordenschat
      2. woorden uit reeds geziene uitbreidingswoordenschat
      3. woorden die nog niet voorkomen in lijsten geziene woordenschat
   2. andere referentiepunten
      1. lijsten uit Common European Framework (A1-C1)
      2. lijsten uit CREA corpus (complexer: enkel met woordvormen)
      3. andere?
   3. bijkomend: frequentie in de tekst
3. bijkomend? classificeren van n-grams?
   1. bijvoorbeeld: 3 of 4-grams met minstens twee lexicale woorden
   2. vergelijken met info uit CREA / collocaties

**Deelproject 3: Verrijken van informatie van lemma’s met internetsearches**

1. info uit
   1. synoniemen / antoniemen
   2. Google Translate
   3. mijn woordenboek
   4. …
2. op basis van
   1. zoekopdracht met lemma
   2. zoekopdracht met verbogen/vervoegde vormen (koppeling met bestaand lexicon)
   3. zoekopdracht met contextuele informatie (bvb. n-gram uit een tekst die als bron dient) (misschien beter voor vertaalsuggesties)

**Deelproject 4: Uitvoeren van (verrijkte) lemma’s naar verschillende didactische applicaties**

1. Voorbeeld 1: uit de analyse in 2 komt een lijst met reeds geziene woordenschat op niveau B1. Voor studenten B2 kan het toch nuttig zijn om deze woordenschat actief op te frissen, bijvoorbeeld door het geven van een vertaling of synoniemen / antoniemen, al dan niet aangevuld met bepaalde tips zoals de eerste letter / alle letters in random volgorde, geven van de context uit de tekst met de eerste letter
2. Voorbeeld 2: (opnieuw voor reeds geziene woordenschat). De lemma’s worden gecheckt tegenover een databank met andere info die opgesteld werd in andere projecten, bijvoorbeeld lijsten met frequente fouten, moeilijke afleidingen substantief/adjectief, frequente collocaties
3. Voorbeeld 3: nieuwe woordenschat:
   1. mogelijkheden bekijken om nieuwe woordenschat in een tekst te visualiseren
      1. printen van zinnen met nieuw woorden
      2. tekst met indicatie van nieuwe woorden (visualisering?)
         1. bvb tekst en onderaan de nieuwe lemma’s in de volgorde zoals ze voorkomen in de tekst of in alfabetische volgorde
   2. weergeven van verrijkte informatie
      1. synoniemen, antoniemen
      2. vertaalsuggesties