Entwicklung linguistischer Ressourcen

Stefania Racioppa
DFKI GmbH

MT, MAT und TM: LOGOS

- Separate Lexika
  - table = Tisch (General)
    table = Tabelle (Data processing)
- Semantische und morphologische Analyse
  - human, animate, etc.
    software, program, etc.
  - Noun (gender, case, number…)
    Verb (transitive, intransitive…)
- Semantische Regeln
  - run(vt) N(text, experiment, etc.) = durchführen N
    run(vt) N(software, programm, etc.) = ausführen N
    run(vt) N(business, firm, corporation, etc.) = leiten N

MT, MAT und TM: Aufgaben

- Textanalyse
- Pre-editing
- Probeübersetzung
- Erweiterung Datenbanken
  - Lexikon
  - Semantische Regeln
- Endübersetzung
- Post-editing

Spell check: Primus

- Identifikation
  - Stamm, erlaubte Endungen
  - Silbentrennung
- Korrektur, Stilkontrolle
  - Autom. Korrektur typischer Fehler (Lybien → Libyen)
  - Konsistentes sprachliches Erscheinungsbild
- Rechtschreibkonverter
  - herkömmlich → neu mit Varianten
  - Sprachregionen (D, CH, A)
- verschiedene Prüfstile
  - Agenturenvorzug (dpa), Duden-Empfehlung

Spell check: Aufgaben

- Erweiterung Rechtschreib-Lexikon (D, E, F, I, S)
  - Neologismen
  - Kunden- bzw. Fachterminologie
- Entwicklung neuer Kodiermodelle
  - Fremdwörter, Fachtermini …
  - Rechtschreibvarianten (D)
- Koordination, Qualitätssicherung
- Programmhilfen, Doku, Schulungsunterlagen
- Kundenbetreuung, Schulungen

Information retrieval: IDX

- Phase 0: Grundformermittlung, Normalisierung
  - Häuser → Haus; Fluß → Fluss …
- Phase M: Mehrwortbegriffe
- Phase B: Ermittlung getilgter Teilwörter
  - Haus- und Hofwirtschaft → Hauswirtschaft, Hofwirtschaft
- Phase 1: Strukturanalyse, Stoppwörter
- Phase 2: Derivation, Dekomposition
  - Personengesellschaft → Person, Gesellschaft
- Phase 3: Aufbau Indexierungsergebnis
- Phase G: Ausgabe Schlagwörter
- Phase T (opt.): Übersetzung Schlagwörter

Information retrieval: Aufgaben

- Rechtschreib-Lexikon
  - Identifikation, Korrektur, Varianten
- Relationenwörterbuch
  - Dekomposition (Deutsch)
  - Synonyme, Derivationen
  - Thesaurus (Erläuterungen, sachverw. Wörter)
  - Übersetzung auf Wortbasis (D, E, F, I, S)
  - ...

Text analysis: SProUT

- Multilinguale NER
  - ENAMEX: Personen, Firmen, Geographika
  - TIMEX: Datum, Zeit
  - NUMEX: Prozentzahlen, Währungsausdrücke
- Einfache Beziehungen zwischen NE
  - „Theodore Roosevelt, President of the U.S.A.“
- Domainspezifische Textanalyse
  - Soccer: Eventketten

Text analysis: Aufgaben

- Erweiterung linguistischer Ressourcen
  - Gazetteer
  - Lexikon
- Entwicklung Analyseregeln
  - Allgemein (numex, enamex, …)
  - Kundenspezifische Domänen (z.B. Events)
- Schulungen, Workshops
- Dokumentation

Zusammenfassend…

- Linguistische Ressourcen für:
  - MT, MAT, TM
  - Spell check
  - Information extraction / retrieval
  - Text analysis
- Andere Tätigkeiten:
  - Kundenbetreuung und -beratung
  - Schulungen, Workshops
  - Programmhilfen, Dokumentation

Danke für Ihre Aufmerksamkeit!