STTS & CLARIN-D

Kathrin Beck

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

CLARIN-D Zentren

- www.clarin-d.de

CLARIN-D – Ziele

- Aufbau einer integrierten, interoperablen und skalierbaren Forschungsinfrastruktur über einen Verbund von Zentren
- Bereitstellung von linguistischen Daten, Werkzeugen und Diensten
- Zielgruppe: WissenschaftlerInnen in den Geistes- und Sozialwissenschaften

Standards

- Verwendung eines passenden **Standards**
- Verwendung von Standards erleichtert die Verwendung und Interoperabilität von Daten und Ressourcen
- Anpassung eines Standards an persönliche Bedürfnisse, möglichst rückführbar auf den Standard
- Ausführliche **Dokumentation** und **Metadaten**, v.a. der Abweichungen vom Standard

ISOcat – Data Category Registry

- ISOcat (http://www.isocat.org): Verzeichnis von Datenkategorien; Referenz-Implementierung von ISO 12620:2009 (ISO/TC37/SC3), einem ISO-Standard zu „Systeme für die Verwaltung von Terminologie, Wissen und Content“
- Datenkategorie (DC) im ISOcat-Kontext: elementarer Deskriptor in einer linguistischen Struktur oder einem Annotationsschema.
- Spezifikationen bestehen aus drei Teilen:
  - Administrativer Teil: Verwaltung und Identifikation der DC
  - Deskriptiver Teil: Dokumentation in verschiedenen Arbeitssprachen
  - Linguistischer Teil: Beschreibung der konzeptuellen Domäne.

ISOcat

CLARIN-D

- Normalisierte Definitionen, z.B. von:
  - Annotationsebenen von linguistischen Daten
  - Tagsets
  - Input und Output-Formaten von linguistischen Werkzeugen
  - ...

ISOcat

- Verwendung von ISOcat:
  - CMDI Metadaten
  - Annotationsstandards MAF, LAF,...
  - WebLicht Webservice zu ISOcat
  - ...

ISOcat

- Nutzen:
  - Interoperabilität von Werkzeugen
  - Suche nach relevanten Korpora und Tagsets
  - Mapping/Interoperabilität verschiedener Tagsets (RELcat)
  - Vergleichbarkeit von Daten
  - Korpusübergreifende Suchanfragen
  - Standardisierte, einheitliche Definitionen zur Wissenskonservierung

STTS & ISOcat

- Einträge der POS-Tags
- Definitionen notwendig über die Guidelines hinaus
- Experten notwendig für die verschiedenen Verwendungen der Tags

Datenkategorien

- Definition der Wortarten mit zugewiesenem Tag, z.B. „Appellativa“, „Eigenname“ mit den zugehörigen Tags
- Definition der übergeordneten Konzepte, z.B. „Nomen“
- Separate Definitionen für verschieden verwendete Wortarten, z.B. „ADJD“ vs. „ADV“:
  - TIGERCorpus: Es ist [ADJD wirklich] schwer
  - TüBa-D/Z: Die sind [ADV wirklich] verdächtig

STTS-„Dialekte“

- Dokumentation aller STTS-„Dialekte“:
  - Abweichungen von Tags:
    - Umbenennungen: PAV → „PROAV“ („Pronominaladverb“) in TIGERCorpus, „PROP“ („pronominale Form einer Präpositionalphrase“) in TüBa-D/Z
    - neu hinzugekommene Tags: z.B. „BS“ in TüBa-D/S
    - weggelassene Tags: z.B. „PIDAT“ in TIGERCorpus
  - Abweichende Verwendung von Tags:
    - „PIAT“ in TIGERCorpus auch für „attribuierende Indefinitpronomen mit Determinierer“ verwendet
    - „ADJD“, „ADV“ in TüBa-D/Z vs. TIGERCorpus

CGN (Corpus Gesproken Nederlands)

- Komplexes Tagset: 316 Tags
- Optimiert für automatisches Tagging ohne weitere Annotationsebenen
- Beispiel: Pronomen „hijzelf“ (himself)
  VNW (pers,pron,nomin,nadr,3m,ev,masc)

CGN (Corpus Gesproken Nederlands)

CLARIN-D

Beispiel: Pronomen „hijzelf“ (himself)

VNW (pers,pron,nomin,nadr,3m,ev,masc):

- VNW DC-4951 → pronoun
- pers DC-4984 → personal
- pron DC-4978 → pronoun
- nomin DC-4941 → nominal
- nadr DC-5008 → stress
- 3m DC-5003 → person 3m
- ev DC-4918 → singular
- masc DC-4930 → masculine

ISOcat „VNW“ screenshot

CLARIN-D

#	Name	Version	Administration stat	Registration status	Check	Type	Owned by	Scope
4951	**pronoun**	1:0	private	private	✓	simple	Schuurman, Ineke	private

pronoun - 1:0

Key	4951
PID	http://www.isocat.org/datcat/DC-4951
Type	simple
Owner	Schuurman, Ineke
Scope	private

1. Administration Information Section

1.1 Administration Record	
Identifier	pronoun
Version	1:0
Registration Status	private
Administration Status	private
Justification	needed for CGN (van Eynde 2004)
Origin	CGN
1.1.1 Creation	
Creation Date	2012-02-27
Change Description	Created for the CGN tagset.

2. Description Section

Profile	Morphosyntax
2.1 Data Element Name Section	
Data Element Name	VNW
Source	CGN
2.2 Data Element Name Section	
Data Element Name	PRON
Source	CGN
[−] **2.3 English Language Section**	
Language	English (en)
2.3.1 Name Section	
Name	pronoun
Name Status	admitted name
2.3.2 Definition Section	
Definition	word referring to an element in a text or outside a text, i.e. in the world outside. There exist, however, a few non-referring pronouns.
Source	CGN
2.3.3 Example Section	
Example	non-referring: it is raining
Source	CGN
2.3.4 Note Section	
Note	There exist, however, a few non-referring pronouns.
[+] **2.4 Dutch Language Section**	

CGN Tag-Komponente „VNW“

Key: 4951

PID http://www.isocat.org/datcat/DC-4951

Administrative Information Section

Identifier: pronoun

Justification: needed for CGN (van Eynde 2004)

Origin: CGN

CGN Tag-Komponente „VNW“

Description Section

Profile: Morphosyntax

Data Element Name: VNW
Source: CGN

Name: pronoun

Definition: word referring to an element in a text or outside a text, i.e. in the world outside. There exist, however, a few non-referring pronouns.

Example: non-referring: It is raining

Note: There exist, however, a few non-referring pronouns.

Fragen?

Interesse am ISOcat-Standardisierungs-Team?