

LEXUS a flexibele web gebaseerde lexicon applicatie

Marc Kemps-Snijders

www.mpi.nl/lexus

lexus@mpi.nl





- Achtergronden
- LMF
 - Core model
 - Componenten en data categorieen
 - Extensie mechanisme
 - Relaties
- Demonstratie



Achtergrond – Situatie bij MPI

7 DOBES teams en 12 verschillende lexica(structuren, doelen)

Tuvan orthography
Tuvan appendix
German orthography
Russian orthography
Russian appendix
Xakas orthography
Tofa orthography

eenvoudig spreadsheet

stem orthography
sense *
lexical sub-entry *

Iets meer complex incl 1:N relaties

sense nr
sense
gram cat
gram subcat
Engl Transl
example *

orthography
Engl. Transl
[T pr] nr

headword
citation form
homograph no
phonetic form

entry-type = [stem idiom lexical word]
head
outer-body-L*

inner-body-L
grammar

sense number
variety
meaning
etymology
table
example*
comment*
picture/photo*
housekeeping*

gloss
word-level-gloss
reversal
definition
encyclopedic info
scientific name
semantic domain
semantic index
thesaurus
semantic relation*
cross-ref*

Klein deel van een complexe lexicon structuur.
Op top niveau 4 verschillende entry types (slechts een is afgebeeld)



Achtergrond - Problemen met lexica

- we hebben een lexicon archief representatie formaat gebaseerd op XML nodig.
- we moeten een archief exploitatie framework bouwen.
- echter, we ontvangen lexica met
 - verschillende character encodings
 - in allerlei formaten (versch. XML, SBX, CHAT, zelfs Word)
 - met verschillende structuren
 - met verschillende terminologieën (attributen, waarden)
- Hoe doen we cross-lexica searches?
- Hoe doen we merging, linking and vergelijking?
- Hoe lossen we lexicon-corpus interactie op?
- etc.



- ISO TC37/SC4 houdt zich bezig met standaardisatie op het gebied van LR Management

- centraal staat **data category registry**

- in essentie een platte lijst van linguïstische concepten
- bevat is_a relations die onderdeel uitmaken van de concept definitie

“transitive_verb” is_a “verb”

- met heldere definities en conceptueel domein(waarde bereik)
- aanvraag voor vullen van DCR (Metadata, morfologie, syntax, ...)

- zoekt naar **abstracte modellen** (frameworks)

- voor lexica
- voor annotatie structuren
- voor semantische annotaties
- voor syntactische annotaties
- ...

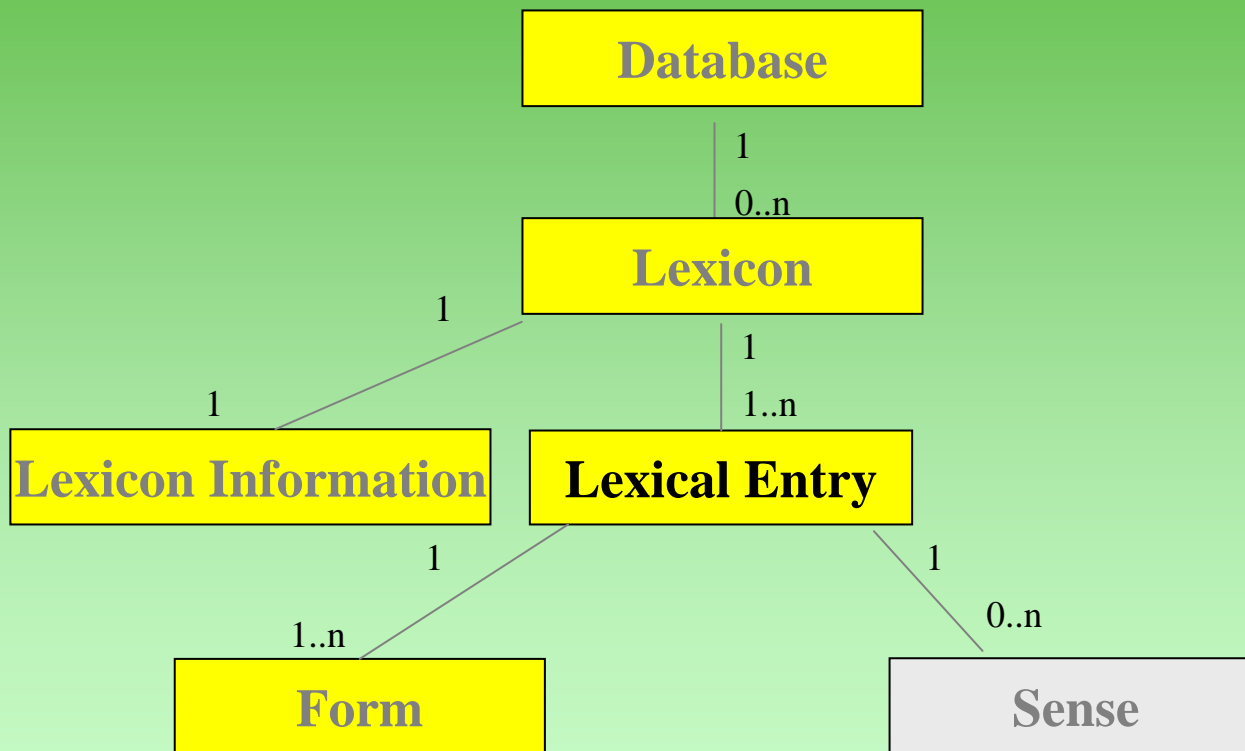


De doelstellingen van LMF zijn

- Een gemeenschappelijk model creëren voor creatie
- En gebruik van zeer grootschalige lexicon bronnen
- De uitwisseling van data in en tussen deze bronnen te beheren, en
- Samenvoegen van grote hoeveelheden kleinere elektronische bronnen tot grote veelomvattende bronnen.
- Het uiteindelijke doel van LMF is een modulaire structuur te creëren om tot ware content interoperabiliteit tussen alle aspecten van lexicon bronnen te komen.

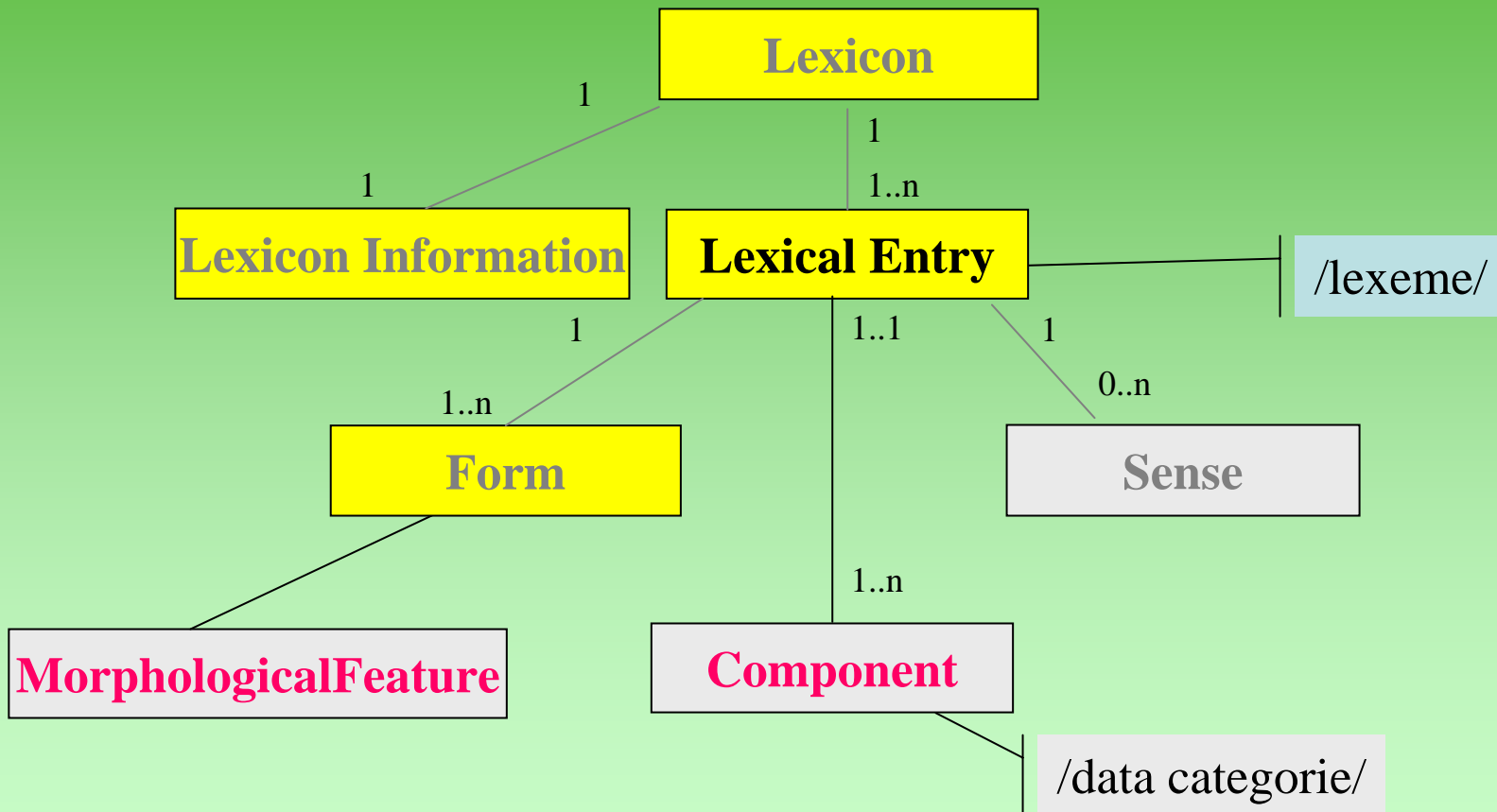


LMF Core Model



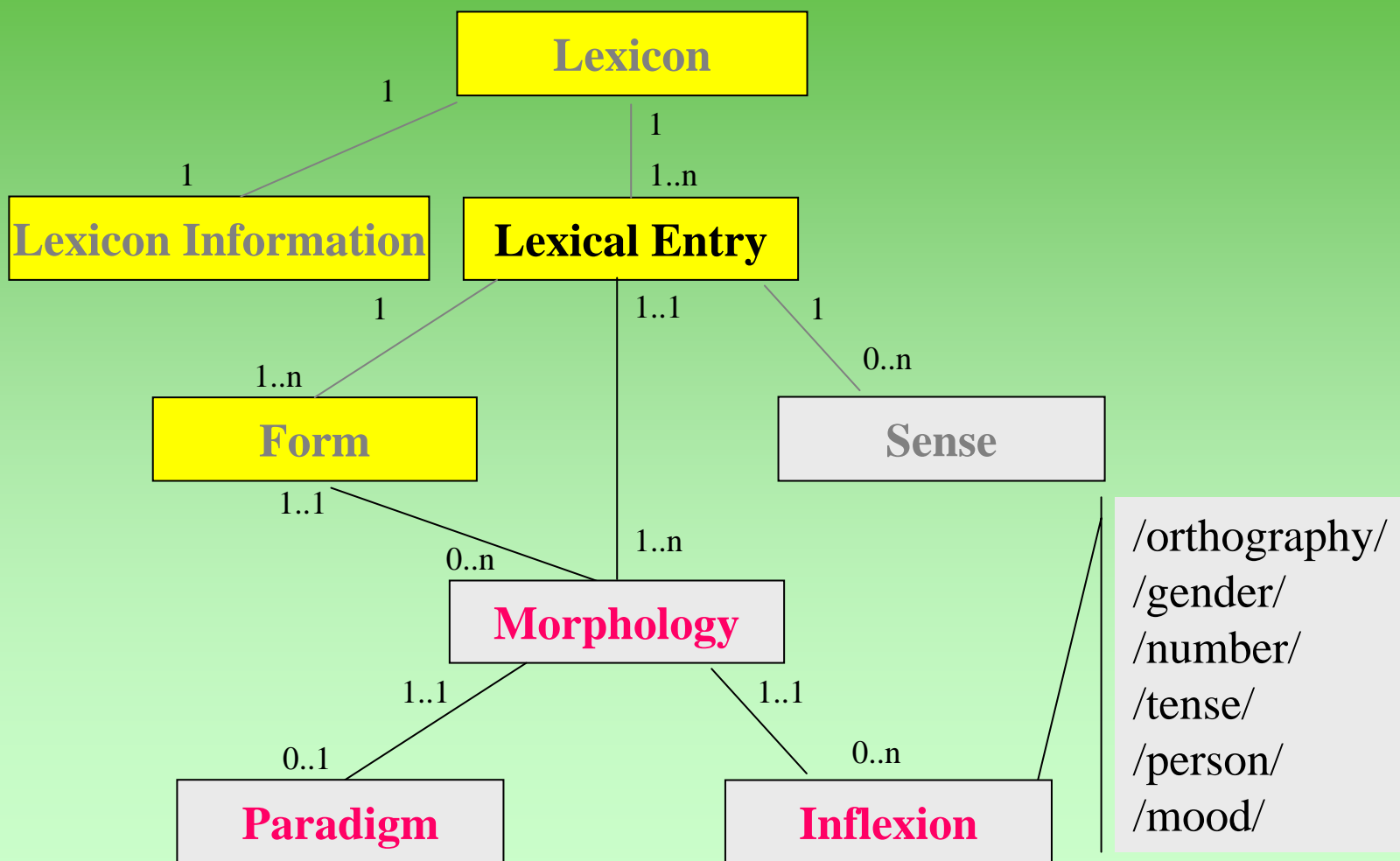


LMF – componenten en data categorieën





LMF Core Model – extensie mechanisme



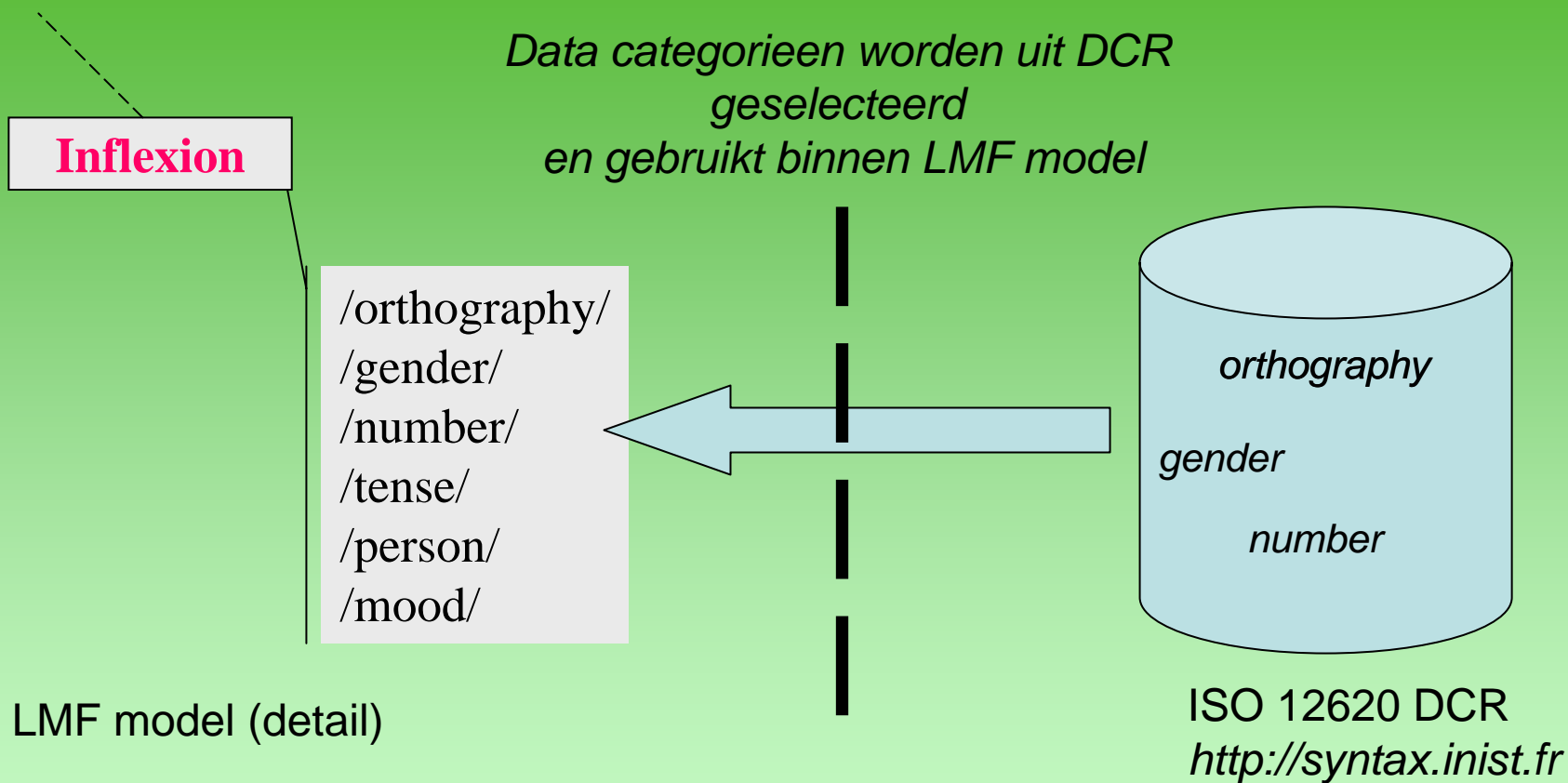


Data categorie selectie

- Grote variatie in attribuut naamgeving en waarden definitie levert problemen op voor interoperabiliteit.
 - Bijv. Het concept noun van worden gemarkeerd als 'n', 'no' of noun.
- Een Data Category Registry bevat een lijst van gestandaardiseerde concepten, beschrijvingen en waarden bereik (ISO 12620, Shoebox MDF, Gold).
- Voor bestaande lexica is het mogelijk unificatie te bereiken door attribuut namen en waarden af te beelden op overeenkomstige data categorieën in de DCR en op te slaan in een apart profiel (bijv. Field tool van E-MELD of OntoELAN van WayneState University)
- Nieuwe lexica kunnen direct naar een DCR verwijzen. (LEXUS)



LMF en Data Category Registry



Uitwisseling vindt plaats via DCR service interface.
(LIRICS project)



DCR interactie (screen shots)

Data Category Repository:

Search: partOfSpeech

Fields: Identifier

Profile: All

All

Private

CategoryAdr

MetaModel

Level

Terminology

LanguageDe

MorphoSynt

Semantic

Personal

Dialog

Data category information partOfSpeech

Administration Identification

Identifier: partOfSpeech	Creation date: 2004-07-09 12620-2:2003; 12620-3:2003
Version: 0.0.0	
Registration authority: Private	Last change date: 2000-09-27 no change description found
Administration status: Private	
Origin: ?	
	Effective date: 2001-09-11

ExplanatoryComment
ISO 12620A-020201

Description

Profile: Terminology



Relaties

- Het moet mogelijk zijn relaties aan te geven tussen verschillende elementen in een lexicon structuur.
 - Het gebruik van relaties wordt in verschillende LMF extensies gesuggereerd.
- Relaties kunnen door gebruikers gedefinieerd worden.
- Gedefinieerde relaties types moeten herbruikbaar zijn.
- Het gebruik van relatie typen moet beperkt kunnen worden tot geselecteerde elementen.
- Het moet mogelijk zijn via relaties te navigeren.