Integrated access to diverse linguistic databases with the Typological Database System

Alexis Dimitriadis
Menzo Windhouwer
Adam Saulwick
Rob Goedemans
Tamás Bíró
Kees Hengeveld

NYU, November 2007

Overview (I)

The Typological Database System (TDS) provides integrated access to multiple, independently created typological **databases.**

Users can query the aggregated databases through the system's **web server:**

http://languagelink.let.uu.nl/tds/

The server software is supported by the data integration **process** that the project has developed.

Overview (II)

- The goal of the system is to provide uniform access to a collection of diverse typological databases:
  - Provide an interface that will help users **find** relevant data.
  - Allow users to **interpret** the data they are presented with.

- The system behaves, as much as possible, as a single database.

  Various differences between the component databases must be dealt with.

Linguistic topics

Taxonomy of linguistic topics, based partially on Thomas Edward Payne's book *Describing Morphosyntax* (Payne, Thomas Edward, 1997).

Grammatical categories
Nouns, Verbs, Modifiers, ...

Constituent order typology
Constituent order with main clauses, Constituent order with verb phrases, Constituent order with noun phrases, ...

Noun and noun-phrase operations
Number, Case, Articles, determiners, demonstratives, ...

Predicate nominals, existentials, possessives, etc.
Predicate nominals, Predicate adjectives (attributive clauses), Predicate locatives, ...

Grammatical relations
Agreement, Functional groupings for S, A, and P, "Syntactic" ergativity

Voice and valence adjusting operations
Reflexives; valence decreasing operations, Valence and predicate calculus, Valence increasing operations

Verb and verb-phrase operations
Nominalizations, Tense/aspect/mode, Location/direction, ...

Pragmatically marked structures
Pragmatic statuses; focus, contrast, and "topicalization", Negation, Non-declarative speech acts

Parts-of-speech system
Differentiated parts-of-speech system, Flexible parts-of-speech system, Rigid parts-of-speech system

Clause combinations,

Discourse analysis

Continuity (cohesion) and

Presentation outline

- Overview of the TDS
- Managing differences between databases
- The TDS user interface
- The component databases

Next:

- Overview of the TDS
- Managing differences between databases
- The TDS user interface
- The component databases

Superficial differences

- Different notational conventions

  e.g. glossing labels, field and variable names, description language

- Different design choices
  - There are many ways to organize information into tables and attributes

- Different software platforms

- Different types of content
  - “Analytical” variables which characterize a language as a whole
  - Annotated sentences with glosses, translations, and descriptive parameters
  - Multiple constructions per language

Contentful differences

- Different theoretical commitments influence:
  - Selection of what is recorded as “data”, and decisions on what factors to control for
  - Criteria and categories to be described
  - Associated terminology
- These differences are deliberate choices;
  If researchers don’t agree on a single analysis, they cannot be resolved.

The TDS approach

- Resolve superficial differences.
- Respect and highlight the theoretical commitments of each database, taking care to preserve the integrity and validity of the data.

Basic Word Order appears in:
Language → Linguistic phenomenon → Clause-level constructions → Basic Word Order

Basic Word Order

add group

Information concerning the canonical order of constituents within the clause
Related concepts.

Basic Word Order combinations

add

Subject, object and verb occur in a certain order, which indicates the basic word order of a language.
Related concepts.

Show the 5 values.
Data source is: Typological Database Nijmegen

Basic Word Order of Clause

add

The basic word order of the clause. Hengeveld, Rijkhoff & Siewierska (2004:542): This is a classification of 'clausal word order in terms of the location of predicates, rather than of verbs, relative to their arguments'. This is based on the order obtaining in 'main, positive, declarative clauses with two overt referential phrases'. 'The major criterion for assigning a basic order is statistical frequency. In languages exhibiting considerable word order variation, we assign a unique basic order only if one of the word order patterns is at least twice as common as any other order, following Dryer (1997).'

Related concepts.

Show the 6 values.
Data source is: Typological Database Amsterdam

Fixed subject predicate order

add

Morphological coding of deviant subject predicate order

add

How databases are integrated (I)

- A dump of the database is made available to the TDS.
- TDS developers define an import schema, which situates the contents of the database in the global hierarchy of the TDS.
- The data undergoes some transformations for uniformity; e.g., **1/0** and **true/false** become **yes/no.**
- Theoretically salient differences are preserved and documented (not removed!)
- The creators of the database are asked to clarify definitions and check the results.

How databases are integrated (II)

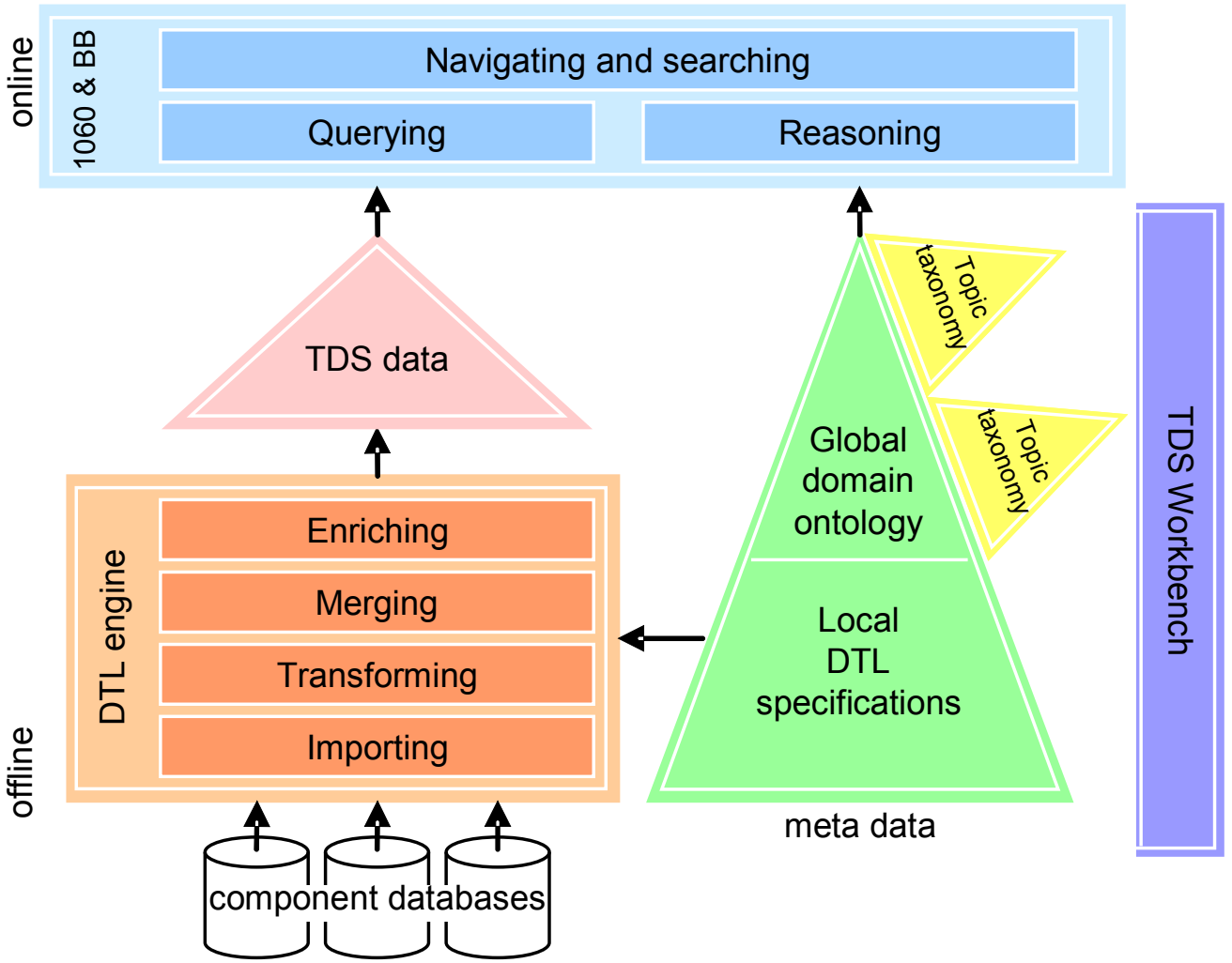
- The import schema is encoded as a combination of

  (a) modular, database-specific documentation and

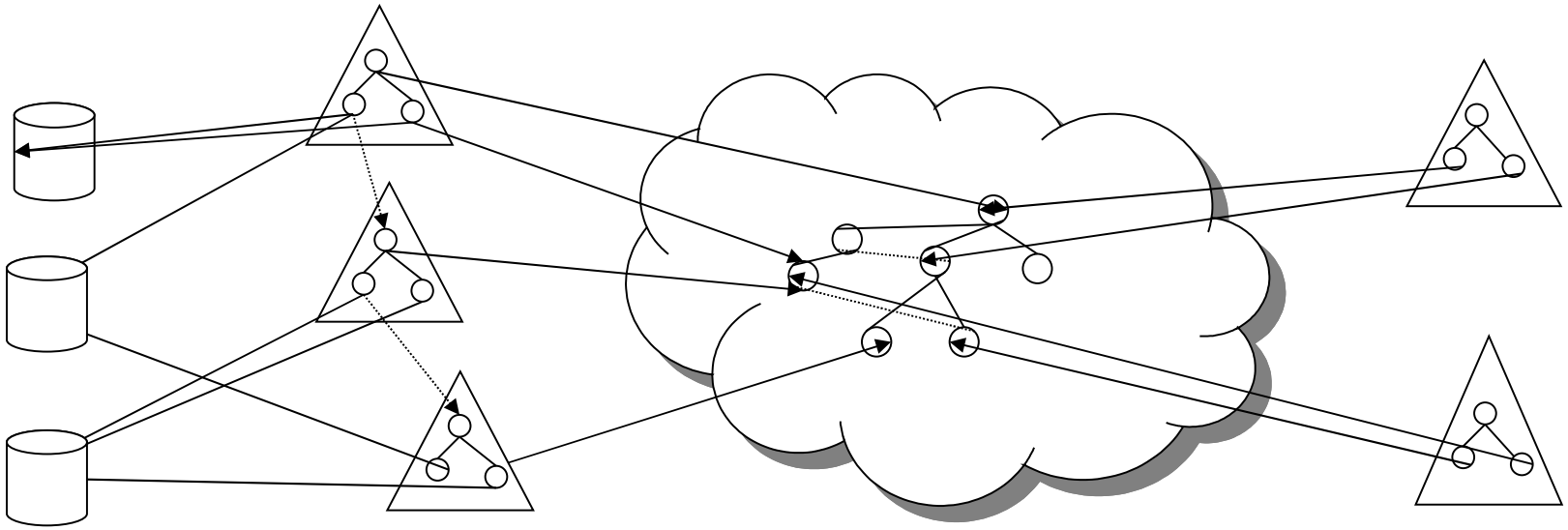
  - pointers into a global ontology of linguistic Concepts

- Data are unified on the basis of ISO language codes (SIL codes)
- The information aids the system in data navigation and presentation, and the users in its interpretation
- Updated versions of the databases can be easily re-imported, using the existing schema

TDS system architecture



Metadata architecture



Database schemata
(any DDL)

Local database ontologies
(DTL)

Global linguistic ontology
(OWL)

Topic taxonomies
(SKOS)

Database developer

TDS Knowledge engineer

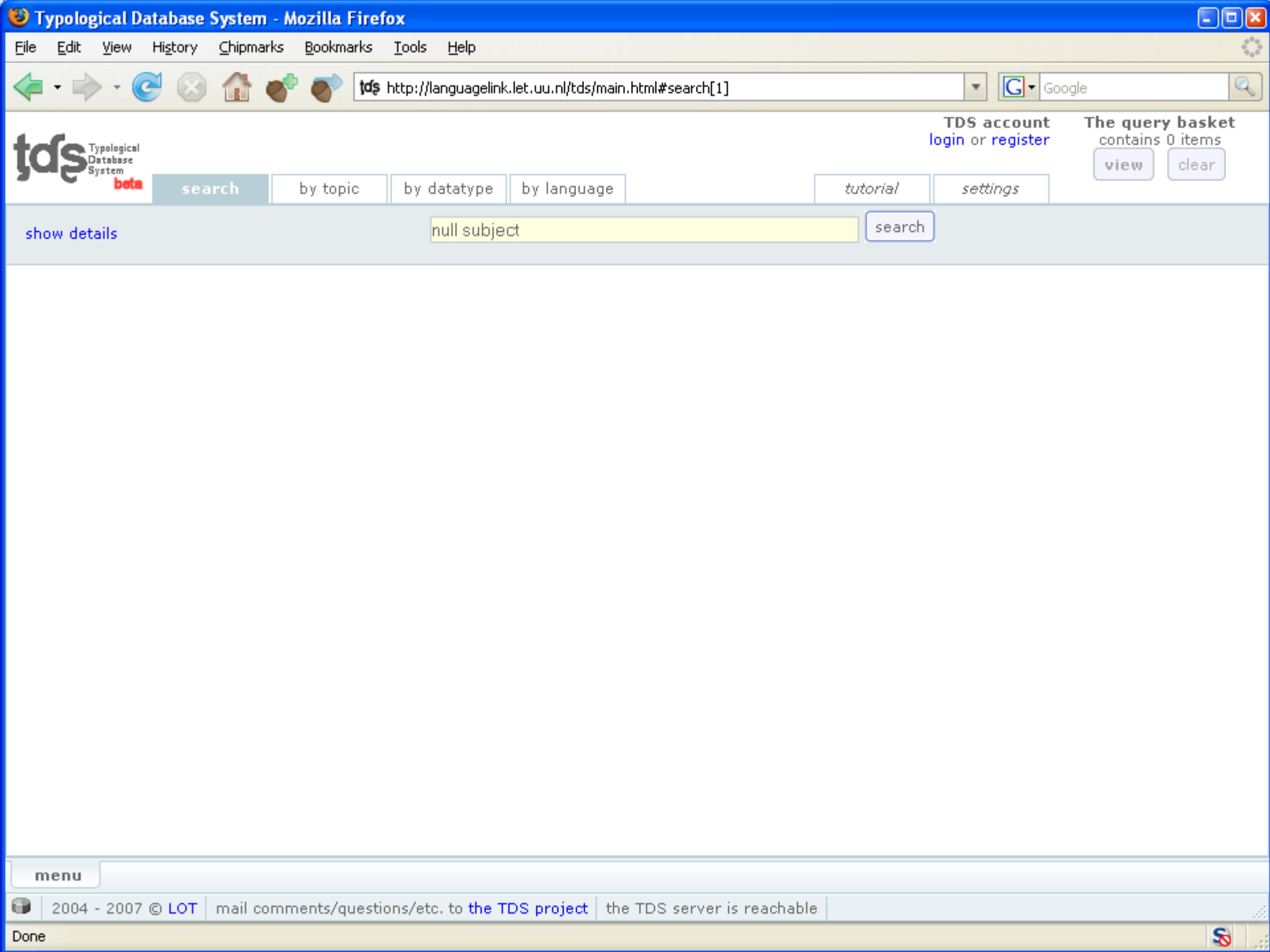
Domain expert

Next:

- Overview of the TDS
- Managing differences between databases
- The TDS user interface
- The component databases

The TDS user interface

- The TDS web interface uses a shopping cart model:
  - Start with collecting interesting items, *i.e.* notions, by searching and/or browsing the meta data;
  - Then look in the basket and fine-tune the query with selection and projection parameters;
  - Execute the query and examine the results in various forms (table, report, statistics, maps, etc.).
- The global ontology is used as a set of structured keywords, boosting related search results, and suggesting alternative searches
- The interface always provides access to the local and global semantic context of a notion or value


Typological Database System - Mozilla Firefox
File Edit View History Chipmarks Bookmarks Tools Help
http://languagelink.let.uu.nl/tds/main.html#search[1]
Google
TDS account
login or register
The query basket
contains 0 items
view
clear
Typological Database System
beta
search
by topic
by datatype
by language
tutorial
settings
show details
null subject
search
menu
2004 - 2007 © LOT
mail comments/questions/etc. to the TDS project
the TDS server is reachable
Done

Description of null subject

Refers to a sentence whose subject is not overtly expressed.

3 matching concepts: covert, intransitive argument (S), transitive argument (A)

There are 328 matches

First Previous **1** 2 3 4 5 Next Last

(There are 14 pages, containing in total 328 items.)

Syntactic features and processes

add group

? Pro-drop

why? add

The language can leave subjects unexpressed with finite verb forms, e.g., Spanish *vivo en Madrid* `I live in Madrid'. This field does not distinguish between "limited" and full pro-drop. For example, Finnish is listed as a pro-drop language but it allows pro-drop in first and second person only.

Related concepts.

Show the 2 values.

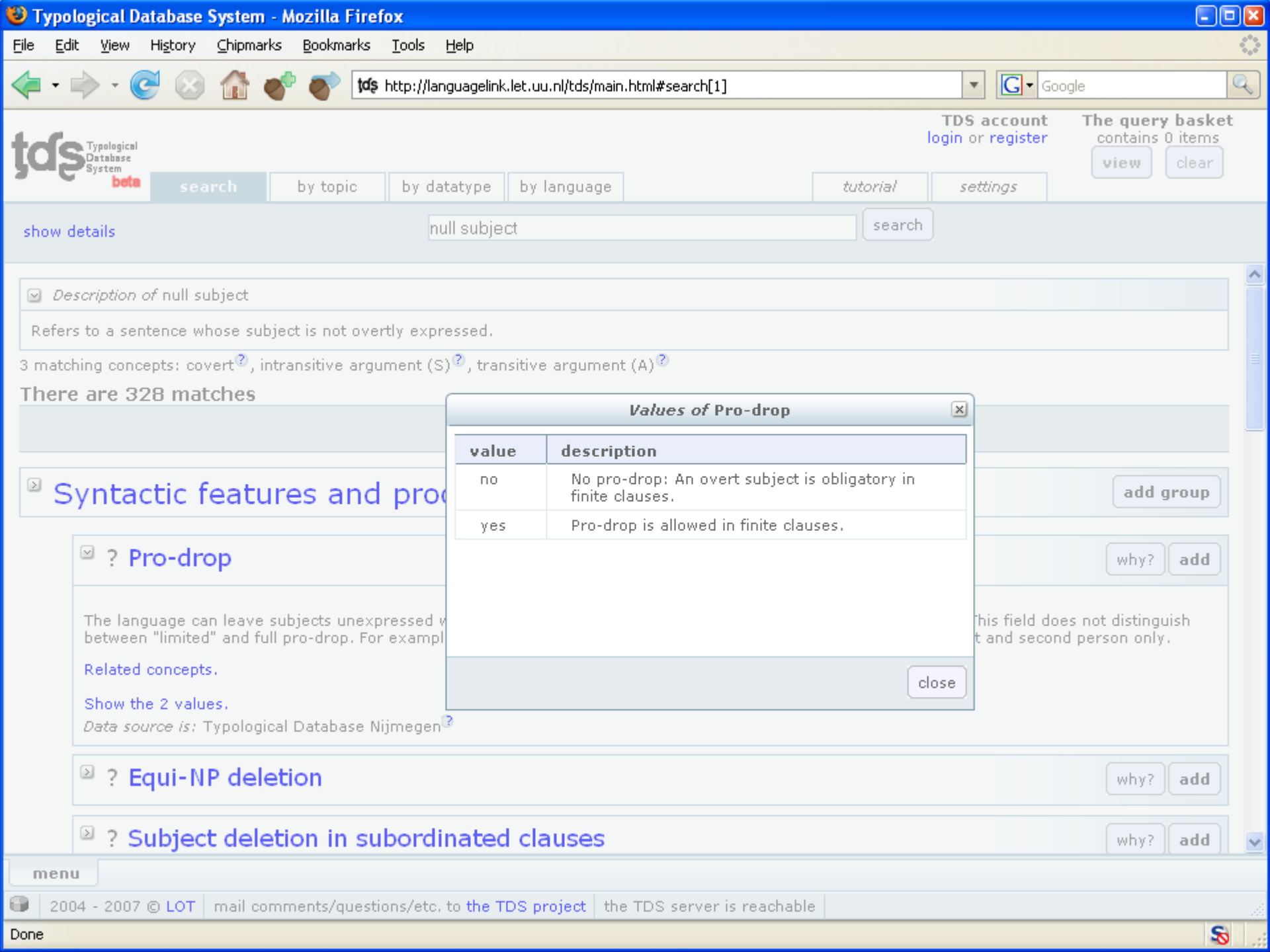
Data source is: Typological Database Nijmegen

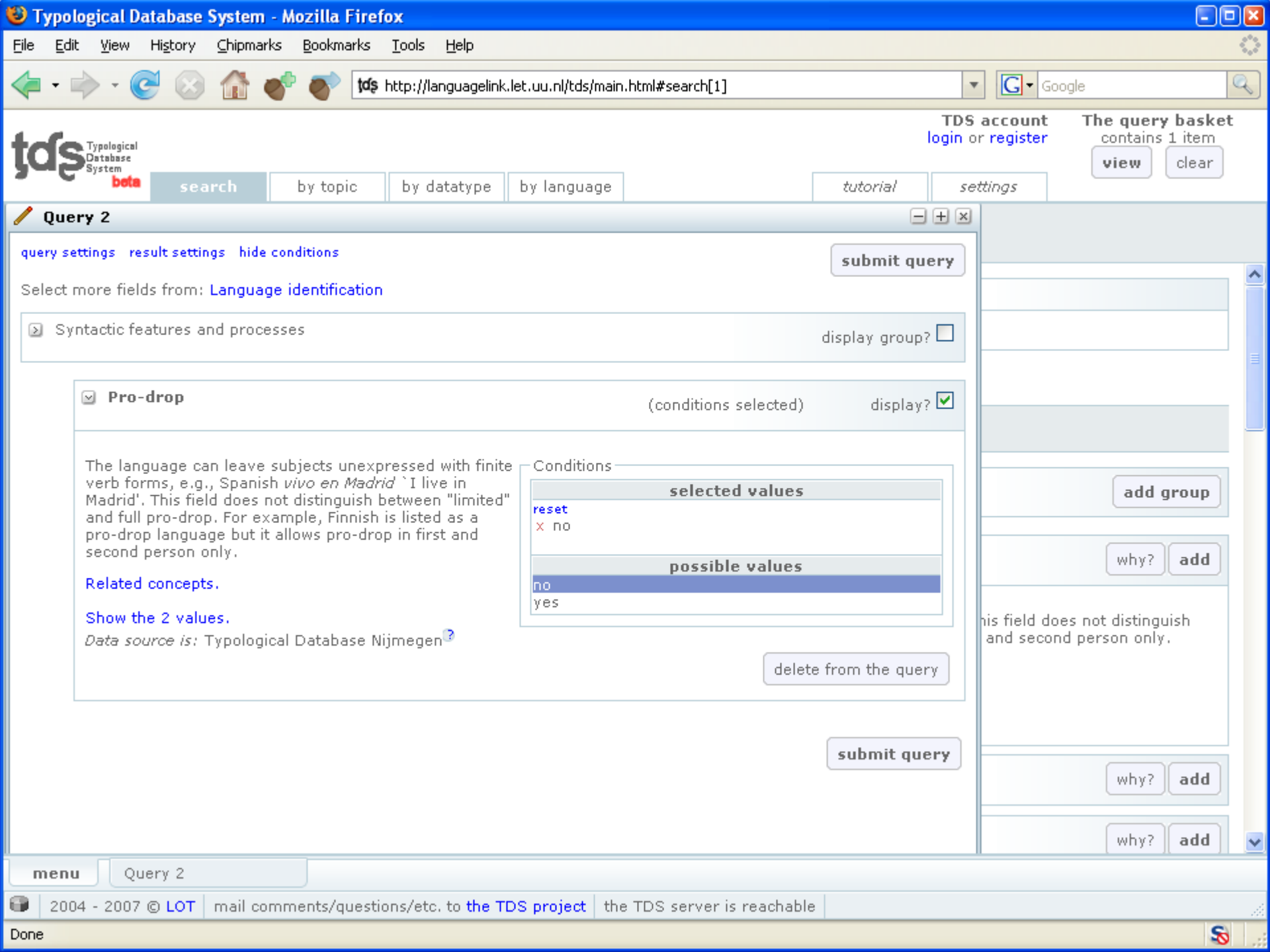
? Equi-NP deletion

why? add

? Subject deletion in subordinated clauses

why? add


Typological Database System - Mozilla Firefox
File Edit View History Chipmarks Bookmarks Tools Help
http://languagelink.let.uu.nl/tds/main.html#search[1]
Google
TDS account
login or register
The query basket
contains 0 items
view
clear
tds Typological Database System beta
search
by topic
by datatype
by language
tutorial
settings
show details
null subject
search
Description of null subject
Refers to a sentence whose subject is not overtly expressed.
3 matching concepts: covert, intransitive argument (S), transitive argument (A)
There are 328 matches
Values of Pro-drop
value
description
no
No pro-drop: An overt subject is obligatory in finite clauses.
yes
Pro-drop is allowed in finite clauses.
close
Syntactic features and proc
add group
? Pro-drop
why?
add
The language can leave subjects unexpressed v
This field does not distinguish between "limited" and full pro-drop. For exampl
t and second person only.
Related concepts.
Show the 2 values.
Data source is: Typological Database Nijmegen
? Equi-NP deletion
why?
add
? Subject deletion in subordinated clauses
why?
add
menu
2004 - 2007 © LOT
mail comments/questions/etc. to the TDS project
the TDS server is reachable
Done


Typological Database System - Mozilla Firefox
File Edit View History Chipmarks Bookmarks Tools Help
http://languagelink.let.uu.nl/tds/main.html#search[1]
Google
tds Typological Database System beta
TDS account
login or register
The query basket
contains 1 item
view
clear
search
by topic
by datatype
by language
tutorial
settings
Query 2
query settings
result settings
hide conditions
submit query
Select more fields from: Language identification
Syntactic features and processes
display group?
Pro-drop
(conditions selected)
display?
The language can leave subjects unexpressed with finite verb forms, e.g., Spanish vivo en Madrid `I live in Madrid'. This field does not distinguish between "limited" and full pro-drop. For example, Finnish is listed as a pro-drop language but it allows pro-drop in first and second person only.
Related concepts.
Show the 2 values.
Data source is: Typological Database Nijmegen
Conditions
selected values
reset
x no
possible values
no
yes
delete from the query
submit query
add group
why?
add
his field does not distinguish
and second person only.
why?
add
why?
add
menu
Query 2
2004 - 2007 © LOT
mail comments/questions/etc. to the TDS project
the TDS server is reachable
Done

Result 1 of query 2

debug info

Show the answer as table . restyle result

Your query resulted in 30 matches for Language.

	Language		Linguistic phenomenon	Clause-level constructions	Syntactic features and processes	Pro-drop
1	Adamawa Fulfulde	browse				no
2	Banda-Banda	browse				no
3	Barclayville Grebo	browse				no
4	Bari	browse				no
5	Biak	browse				no
6	Chamorro	browse				no
7	Dutch	browse				no
8	English	browse				no
9	Ewe	browse				no
10	Farefare	browse				no
11	Fijian	browse				no
12	Igbo	browse				no
13	Ilwana	browse				no
14	Kaingang	browse				no
15	Khasi	browse				no

Typological database entity types

Languages
All linguistic and non-linguistic information about a particular language.
Language identification, LINGUISTIC PHENOMENA, Source information, ...

Universal or cross-linguistic
Collections of information that is not specific to one language or group of languages
Universal phonemic segment pool

Text materials
Text materials (words, morphemes, sentences, etc.)
Example sentences, Examples of some phenomenon

Auxiliary
Non-linguistic information
Interlinear gloss definitions, Bibliography

Next:

- Overview of the TDS
- Managing differences between databases
- The TDS user interface
- The component databases

The component databases (I)

1. Person-Agreement database (A. Siewierska, D. Bakker)
   Person and agreement phenomena. Over 400 languages

2. Typological Database Nijmegen (L. Stassen)
   Word order, predication, case marking, relative clauses, comparatives, possession, coordination, and more.
   Between 140 and 400 languages, depending on topic

3. Typological Database Amsterdam (K. Hengeveld)
   Basic word order and constituent order systems; parts-of-speech systems

The component databases (II)

1. StressTyp (R. Goedemans, H. van der Hulst)
   metrical systems (stress, foot types, extrametricality etc.) for 510 languages

2. SylTyp (H. van der Hulst, R. Goedemans)
   syllable structures

3. UCLA Phonological Segment Inventory (I. Maddieson)
   segment inventories with phonological features for 451 languages

4. Smith‘s Phoneme Inventories (N. Smith)
   Phoneme and lexical tone inventories for 111 languages

The component databases (III)

- Anaphora Typology database (A. Dimitriadis, M. Everaert, E. Reuland, T. Reinhart)
  examples of reflexives with analysis; only a few languages are in the database
- Berlin database of intensifiers and reflexives (V. Gast, D. Hole, E. König, P. Siemund, S. Töpper)
  properties and examples for over 100 languages
- Graz database on reduplication (B. Hurch, V. Mattes, O. Konovalova)
  phonology, morphology and semantics of reduplication, with information on productivity and diachrony

Auxiliary resources

- ISO 639-3 language codes
  Three-letter codes (the former Ethnologue/SIL codes)
- Genetic affilation according to the Ethnologue (SIL International)
- Universal Phoneme Positioning Chart
  Table of potential phonemes, derived from UPSID data with additional processing

To be added

- World color survey (P. Kay, B. Berlin, L. Maffi, W.R. Merrifield)
  Summary information on color term systems
- Topic-focus database (E. Aboh, K. Hengeveld)
- Berlin-Utrecht reciprocals survey (M. Everaert, E. König, V. Gast, A. Dimitriadis, C. Emkow, T. Hanke)
  Inventories of reciprocal markers with their properties, examples
- SCALA/Spinoza database (P. Muysken, M. Klamer, S. Musgrave, H. van Halteren)
  Linguistically annotated texts
- GIS coordinates
  Geographic locations for languages in the system (M. Dryer / WALS; G. Segerer for African languages)

Still to come

- More databases
- Geographic coordinates and map display
- Interface improvements
- Performance improvements

Lessons for the SSWEL

- Organize parameters in groups and structure them hierarchically
- Obligatorily anchor each language or variety by its ISO code (SIL code)
- Support alternative perspectives on the same property, and make sure they are presented side by side
- Make parameter documentation an inherent part of the interface design
- Make it easy to define, and extend, lists of possible values (with documentation)– and to reuse them!
- Provide facilities for *finding* parameters of interest

Other remarks

- The database should contain descriptions of 3 “entity types”:
  - Languages
  - Constructions / Phenomena / Morphemes
  - Sentences

http://languagelink.let.uu.nl/tds/

- Steering committee: Martin Everaert (UiL-OTS), Kees Hengeveld (UvA), Roeland van Hout (RU), Pieter Muysken (RU), John Nerbonne (RUG), Peter Wittenburg (MPI)
- Principal developers: Menzo Windhouwer, Alexis Dimitriadis, Adam Saulwick (until August 2006), Rob Goedemans, Tamás Bíró (since October 2006)
- Interns: Eugenie Stapert, Franca Wesseling, Ruth Lind
- The creators of the component databases have contributed the all-important content, as well as expertise and assistance in integrating their databases into the TDS.
- We gratefully acknowledge the financial support of the Netherlands Organization for Scientific Research (NWO).