Natural Language Processing research at IwiSt: A technology-oriented overview

Ulrich Heid and the STCL team

Universität Hildesheim

Research cooperation workshop, 6-11-2013

Overview

- Personnel –
  Expertise in selected domains of NLP
- Main lines of research:
  Methods under development - domains of application
- Technology from ongoing projects:
  - The e-Identity Text Exploration Workbench
  - Tools for guiding users in lexical selection,
    from the SeLA project
- Medium term planning –
  opportunities for cooperation

Who we are

Personnel working in the domain of language technology/computational linguistics

- Dr. Folker CAROLI — Linguistics
- Getrud FAASZ, Ph.D. — African Linguistics, CL
- Dr. Ulrich HEID — Computational Linguistics
- Dipl.-Ling. Ronny JAUCH — Computational Lingusitics
- Fritz KLICHE, M.A. — Computational Linguistics
- Josef Ruppenhofer, Ph.D. — Computational Linguistics
- Luigi SQUILLANTE — Physics, CL (U.Rome I)

Who we are – teaching topics

Expertise in selected domains of NLP and CL

Porgramming: Perl/Python	Jauch, Faaß
Databases	Jauch
Statistics for NLP	Ruppenhofer
Corpus Methods	Faaß, Heid
Syntax, Morphology,	Caroli, Heid
Lexical Semantics	Ruppenhofer
Dialogue Systems	Heid
Machine Translation	Caroli, Heid
(e-)Lexicography	Heid
Sentiment Analysis	Ruppenhofer

Reseach topics in NLP at IwiSt

Status as of autumn 2013

- Main lines of research (and teaching)
  - Sentiment Analysis – Opinion Mining:
    Using corpus-derived linguistic knowledge — Ruppenhofer
    → More details in J. Ruppenhofer's talk
  - Corpus technology: — Faaß, Kliche, Jauch
    Data acquistion – tools for corpus compilation and processing
  - Electronic dictionaries: — Faaß, Jauch
    User-centered models and GUI design
- Other research work (medium term perspective)
  - Terminology extraction from text
  - Resource infrastructures and standardization

Research topics – Overview

Development of methods – tests within applications

Applications → Methods/Tools ↓	Sentiment Analysis	Lexicography, e-Dictionairies	Digital Humanities
Linguistics as a foundation for tools			
Corpus techniques - cp. collection - acquistion of linguistic data			
User interfaces			

Research topics – Overview

Development of methods – tests within applications

Applications → Methods/Tools ↓	Sentiment Analysis	Lexicography, e-Dictionairies	Digital Humanities
Linguistics as a foundation for tools	Lexical Semantics	Lexical data description	Text structure
Corpus techniques - cp. collection - acquistion of linguistic data	 sentiment lexicons	 collocations, terminology	 sampling, extraction of metadata text form of metadata (e.g. date)
User interfaces		user guidance in e-dictionairies	GUI for corpus processing pipelines

Research topics – ongoing Ph.D. work

Related with corpus-based approach and tools

- Corpus compilation, sampling, metadata extraction
  using linguistic and textstructural knowledge, and ML:
  Fritz KLICHE
- Corpus annotation: improvement of large-scale applicability of
  tokenizing, word class tagging, compound analysis etc.:
  Heike STADLER (working at IdS, Mannheim)
- CL-based control of specification documents for
  consistency, unambigious wording, etc.
  Jennifer KRISCH (working at Daimler, Böblingen)
- Corpus-based study of the English of German learners:
  Verena MÖLLER (working at a highschool in Waiblingen)
- Methods and tools
  for multiword extraction from text of inflecting languages (Italian):
  Luigi SQUILLANTE (Co-Supervision with Uni Roma La Sapienza)

Technology for ongiong projects

The e-Identity Text Exploration Workbench

- Project:
  - Headed by political scientist C. Kantner (Stuttgart)
  - Cooperation with CL from Stuttgart and Potsdam
  - Funded by BMBF ("e-Humanities"): 05/12 – 04/2015
- Objectives of the project:
  - Support political scientists in corpus based work: collecting, sampling, annotating, managing, ... text data
  - Support political scientists in finding textual evidence for abstract concepts: *identity concepts evoked*
    * Beyond keyword-based search
    * Identifying opinion holders and their position
    * Mapping variable textual statements onto concepts

Technology from e-Identity

Functions of the Text Exploration Workbench — Kliche

- Input: texts from news archives, e.g. Lexis-Nexis
- Processing steps:
  - Character encoding and conversion to XML
  - Identification of metadata: date, autor, ...
  - Identification of text structural elements:
    header, byline, body, captions, ...
  - Identification of issue cycles:
    how many articles per newspaper on a given topic, per timespan?
  - ...
- Output:
  Homogeneized texts plus metadata:
  - meta-data on the texts: author, date, text elements, ...
  - process metadata: procedures and tools applied

Technology from e-Identity

Ongoing and upcoming work on processing workflows

- First version of Workbench GUI available (ongoing work):
  - includes several tools from the above
  - provides preview of topic analysis (LDA-based)
- Design study for more detailed Workbench GUI:
  - Interactive, "wizard like" tools
  - User decides for properties the texts should have after processing – GUI proposes a pipeline of CL tools:
    Users don't know details of tool functions and interdependencies

Technology from ongoing projects

SeLA - Scientific e-Lexicography for Africa

- Project:
  - Headed by U. Hildesheim
  - Partners at Universities of Pretoria, Stellenbosch, Windhoek, and at Univerisity of South Africa (UNISA)
  - Funded by BMBF, administered by DAAD: 06/2012 - 05/2015
- Technological objectives of the project:
  - Design of new kinds of lexical information tools, taking South African situation as an example
  - Work on sample dictionaries and terminologies for African languages

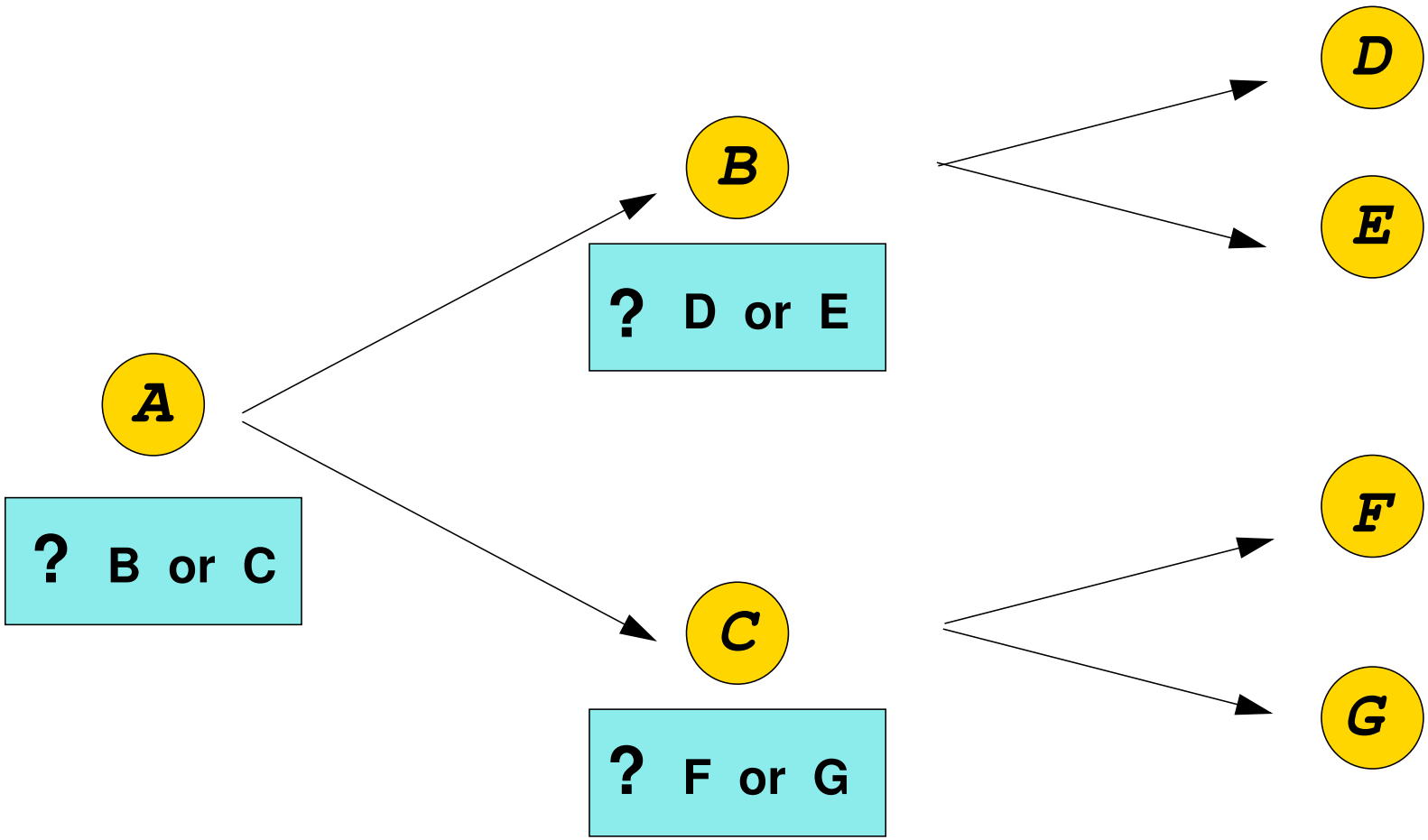
Technology from SeLA

Tools for guiding users in lexical selection

- Why?
  - Many grammatical phenomena of the S.A. Bantu languages show systematic variation:
    * different semantic classes for what is one class in e.g. EN
    * 15 noun classes (like declension classes)
  - Users struggle with, e.g. personal and possessive pronouns
- Approaches:
  (1) Stepwise guidance through a decision tree for lexical selection, see schemata below — Bothma/Prinsloo (Pretoria)
  (2) Guidance in a mono- or bilingual selection task: user-definded amounts of support — Faaß/Bosch (Pretoria)

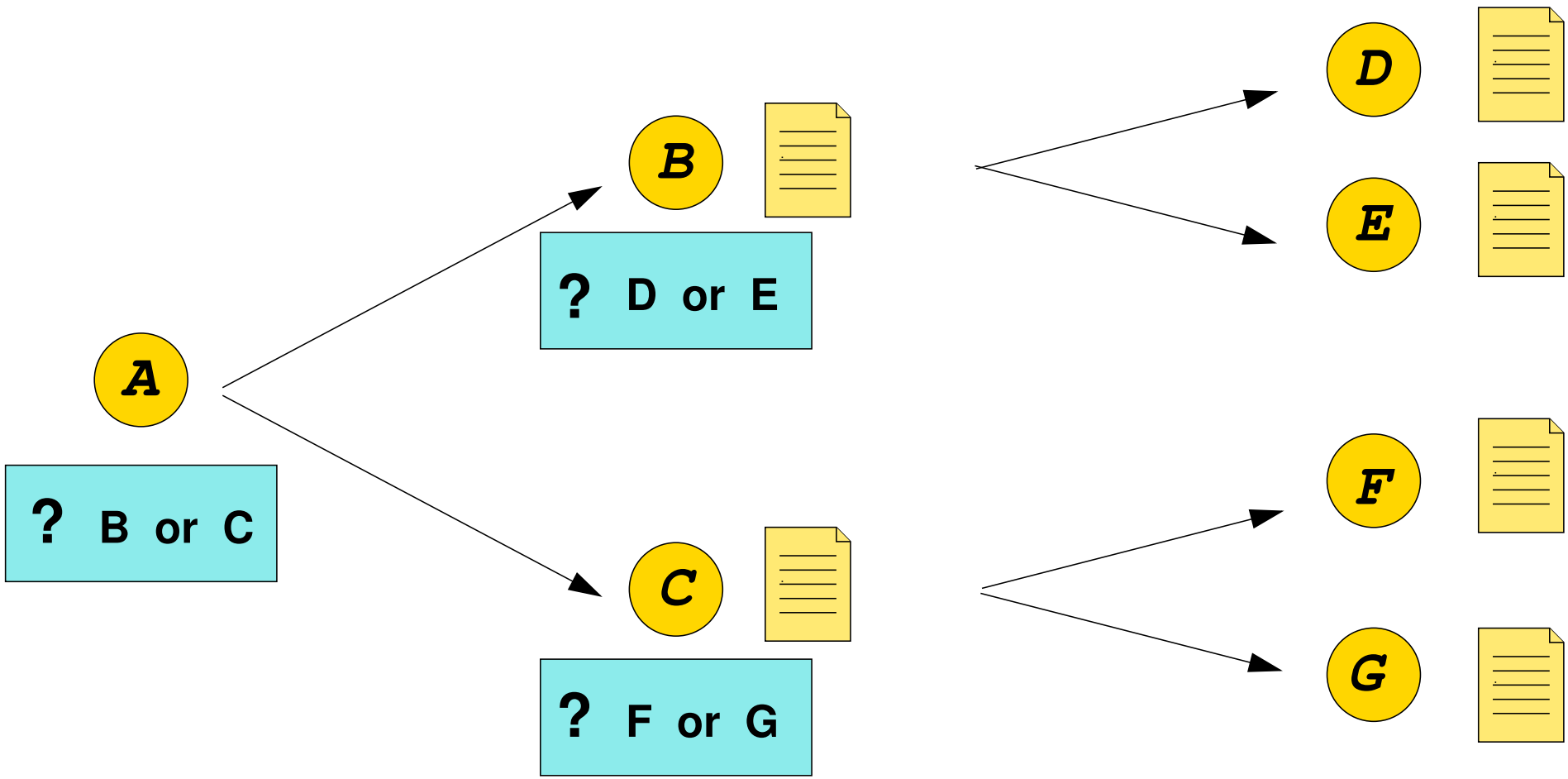
Technology from SeLA

Principles of guidance via decision trees: basic tree



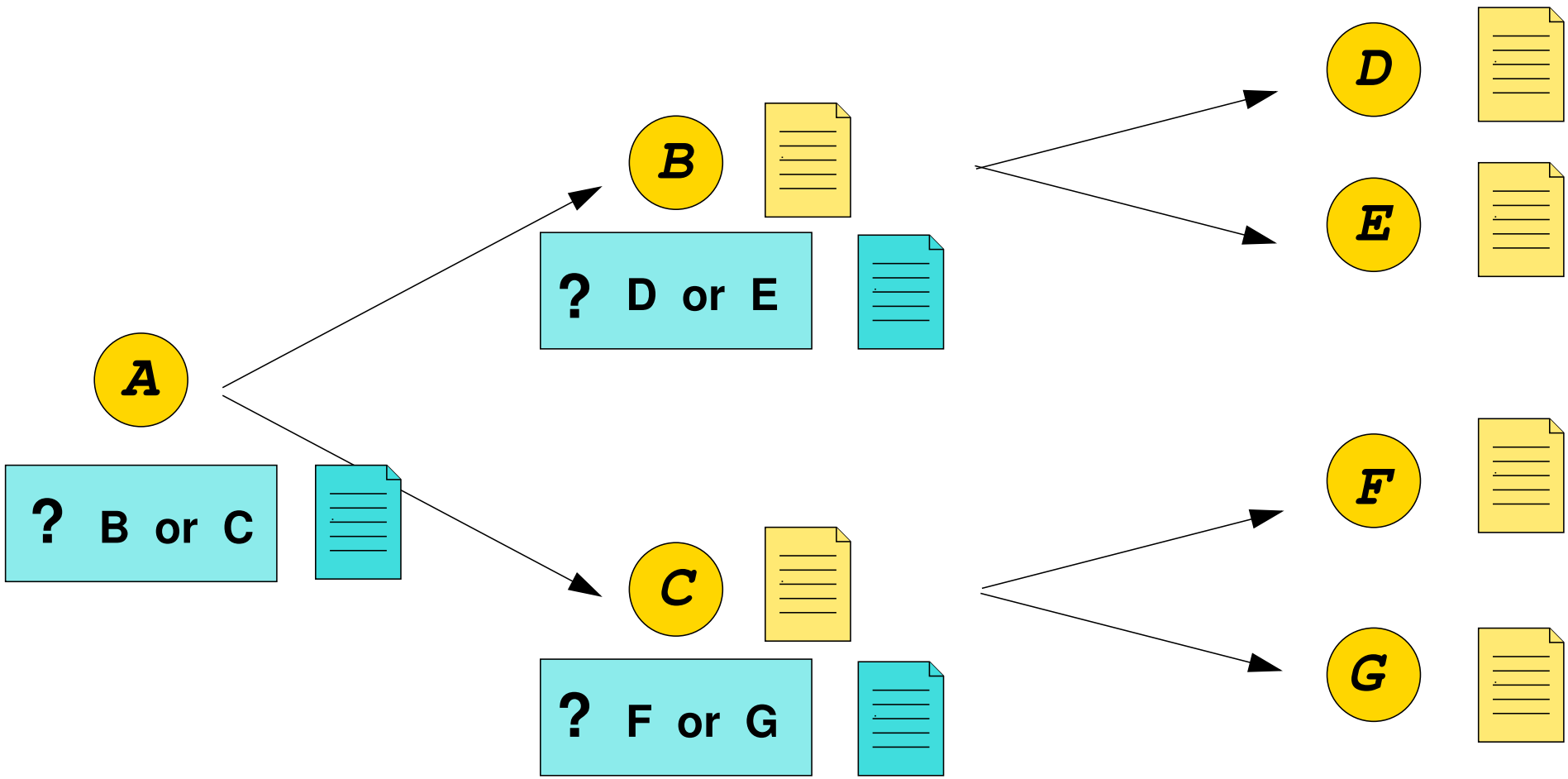
Technology from SeLA

Principles: information on demand about selectionable items



Technology from SeLA

Principles: information on demand at choice points



Technology from SeLA

General approach

- Lexical data in databases — Faaß/Jauch
- Constraints for selection ideally represented separately
- GUIs reachable as Web services via the internet – interaction with lexicon database services — Faaß/Jauch
- Visualization and search tools on the Web — Bothma (Pretoria)

Medium term planning

- More experiments on topics from e-Identity and SeLA:
  - Workflows for digital humanities research
  - GUIs for workflows, GUIs for dictionaries
- Project on sentiment analysis and
  sentiment data mining for EN and DE (submitted)
- Work on terminology extraction:
  Techiques, workflows, GUIs

$\Rightarrow$ Interested in cooperation!