# *Seminar*

# *Lexicography: Projects and Principles*

Dafydd Gibbon

Bielefeld University
Jinan University

23.11.2018

# *Lexicon types – informal overview*

- Word lists (including comparative word lists)
- Thesauri
- Wordnets
- Framenets
- Concordances
- Dictionaries:
  - bilingual and multilingual dictionaries
  - terminologies
  - onomasiologies
  - phrasal idiom lexica
  - Pronunciation dictionaries
- Linguistic inventories:
  - Phonemes
  - Prosody: tones, pitch accents, lexicons of intonation tunes.

… and everybody knows what "the dictionary" is ...

# *Main lexicography Projects since 1989*

Verbmobil – lexicography coordinator (German National)

- multilingual lexicographic model for speech-to-speech translation
- lexicographic infrastructure for German and international research institutes and companies

Ulex – cooperation on language documentation (VWF)

- Ubiquitous comprehensive online dictionary
- Uyghur case study

Uyo Ibibio Dictionary – curriculum development (DAAD)

- bilingual dictionary of a Niger-Congo language

ILEX (Integrated Inheritance Lexicon) – long-term research

- DATR, Inheritance Lexicon Language
  - morphology, prosody, idioms

# *Lexicographic Reports and Publications 1992 - 2007*


Gibbon, Dafydd (1992). ILEX: A Linguistic Approach to Computational Lexica. *Computatio Linguae, Zeitschrift für Dialektologie und Linguistik*, 73:32-53.

Gibbon, Dafydd (1993). Generalised DATR for Flexible Lexical Access: Prolog Specification. VERBMOBIL Report 2. Universität Bielefeld.

Gibbon, Dafydd, Doris Bleiching, Julie Carson-Berndsen, Hagen Langer, Martina Pampel, Matthias Erhard, Christoph Schillo and Markus Vogt. (1994). Bellex3 - Bielefeld Engine for Lattice-to-Lattice Event-parsing. Technischer Report, Universität Bielefeld.

Gibbon, Dafydd (1995). VERBMOBIL Lexicon: Conventions for Spelling and Pronunciation. VERBMOBIL Technisches Dokument 31. Universität Bielefeld.

Gibbon, Dafydd and Ute Ehrlich (1995). Spezifikation für ein VERBMOBIL-Lexikondatenbankkonzept. VERBMOBIL Memo 69. Universtität Bielefeld, Daimler Benz AG, 1995.

Gibbon, Dafydd, Komdedzi Kofi Folikpo, Shu-Chuan Tseng (1996). Prosodic inheritance and phonetic interpretation: lexical tones in Ewegbe. Technical Report, Bieiefeld.



Gibbon, Dafydd, Komdedzi Kofi Folikpo, Shu-Chuan Tseng (1996). Prosodic inheritance and phonetic interpretation: lexical tones in Ewegbe. Technical Report, Bieiefeld.

# *Lexicographic Reports and Publications 1992 - 2007*


Gibbon, Dafydd (2000). Computational Lexicography. In: Frank van Eynde and Dafydd Gibbon, editors, *Lexicon Development for Speech and Language Processing*, Dordrecht: Kluwer Academic Publishers, 1-42.

van Eynde, Frank and Dafydd Gibbon, eds. (2000). *Lexicon Development for Speech and Language Processing*. Dordrecht: Kluwer Academic Publishers.

Gibbon, Dafydd and Harald Lüngen (2000). Speech Lexica and Consistent Multilingual Vocabularies. In: Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer Verlag, 296-307.

Gibbon, Dafydd, Harald Lüngen and Andreas Witt (2000). Enhancing speech corpus resources with multiple lexical tag layers. *Proceedings of LREC 2000*, Athens.

Gibbon, Dafydd and Thorsten Trippel (2000). A multi-view hyperlexicon resource for spoken language system development. *Proceedings of LREC 2000*, Athens, p. 1713-1718.

Gibbon, Dafydd (2002). Prosodic information in an integrated lexi con.*Proceedings of the 1st International Conference on Speech Prosody 2002*. Aix-en-Provence, 335-338.

Nixon, Stephanie M., Malcom R. McNeil, Dafydd Gibbon, Hillel J. Rubinsky, Patrick J. Doyle, Tepanta R. D. Fossett, Grace H. Park. and William D. Hula. The serial position effect and lexical processing during story-retelling in adults with and without aphasia. *Clinical Aphasiology Conference*, Big Cedar, MO, May, 2002

# *Lexicographic Reports and Publications 1992 - 2007*


Gibbon, Dafydd, Thorsten Trippel, Felix Sasaki and Benjamin Hell (2003). Acquiring Lexical Information from Multilevel Temporal Annotations. *Proceedings of Eurospeech 2003*, Geneva.

Gibbon, Dafydd (2004). First steps in corpus building for linguistics and technology, Workshop Proceedings, *First Steps for Language Documentation: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*. LREC 2004, Lisbon.

Gibbon, Dafydd, Catherine Bow, Steven Bird and Baden Hughes (2004). Securing interpretability: the case of Ega language documentation. *Proceedings of LREC 2004*, Lisbon.

Gibbon, Dafydd, Thorsten Trippel and Felix Sasaki (2004). Consistent storage of metadata in inference lexica: the MetaLex approach. *Proceedings of LREC 2004, Lisbon*

Gibbon, Dafydd (2005). Spoken language lexicography: an integrative framework. In: Lew Zybatow, ed. (2005), *Translatologie – Neue Ideen und Ansätze*. Frankfurt, etc.: Peter Lang, Europäischer Verlag der Wissenschaften, 247-289.

Borchardt, Nadine and Dafydd Gibbon (2006). Didactique en documentation de Langues: Aspects de la lexicographie numérique. *Proceedings of West African Linguistics Conference*, Ouidah, Benin, 30.07.2006-06.08.2006.

Gibbon, Dafydd and Nadine Borchardt (2007). Computational lexicography: a training programme for language documentation in West Africa. In: B.M. Mbah and E.E. Mbah, eds. (2007), *Linguistics in History: Essays in honour of P.A. Nwachukwu.* Nsukka, Nigeria: University of Nigeria Press.

# *Main lexicography Projects since 1989*

## Verbmobil

- multilingual lexicographic model for speech-to-speech translation
- Lexicographic infrastructure for German research institutes and companies

## ULex:

- Ubiquitous comprehensive online dictionary
- Uyghur case study

## Uyo Ibibio Dictionary:

- bilingual dictionary of a Niger-Congo language

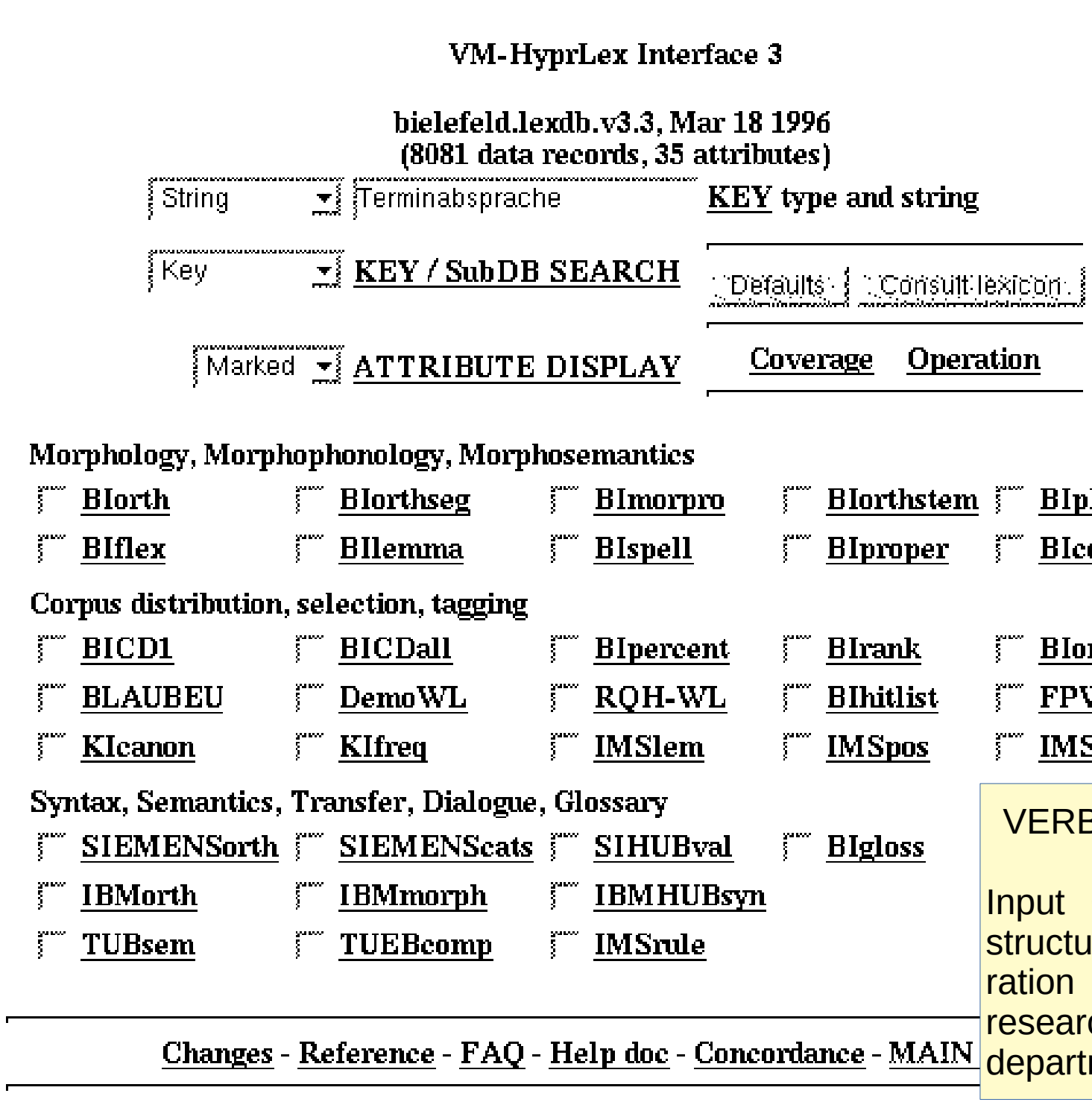
## ILEX (Integrated Inheritance Lexicon):

- DATR, Inheritance Lexicon Language
  - morphology, prosody, idioms

***VERBMOBIL: Speech-to-Speech Translation***

***Objectives:***
***End-to-end speech translation coordination***
***Languages: German, English, Japanese***

***Subproject: Lexicon***
***Lexicographic coordination and integration between international research institutions and companies***




VERBMOBIL ONLINE LEXICON

Input GUI for access to micro-structure for collaborative integration of lexica from different research institutes and research departments of companies.

| *Category* | *Value* |
|---|---|
| BIorth: | Terminabsprache |
| BIorthseg: | Termin#ab#sprach#+e |
| BImorpro: | tE6.m'i:n#?''ap#Spr''a:.x#+@ |
| BIorthstem: | Termin#ab#sprach |
| BIphonstem: | tE6.m'i:n#?''ap#Spr''a:x |
| BIflex: | N,akk,sg,fem;N,dat,sg,fem;N,gen,sg,fem;N,nom,sg,fem |
| BIlemma: | Terminabsprache |
| BIspell: | -- |
| BIproper: | -- |
| BIcompsem: | ObjEreig |
| BICD1: | cd1=2_cd12=7_cd3=2_cd4=3_cd5=1 |
| BICDall: | 15 |
| BIpercent: | 0.00568005% |
| BIrank: | 977 |
| BIortherror: | Termin-Absprache,- |
| BLAUBEU: | -- |
| DemoWL: | demo-wl |
| RQH-WL: | -- |
| BIhitlist: | hit#977=15 |
| FPWL3: | fpwl |
| KIcanon: | tE6m'i:n#Q"ap#Spr"a:x@ |
| KIfreq: | 14 |
| IMSlem: | Terminabsprache |
| IMSpos: | NN |
| IMSfreq: | 8 |
| SIEMENSorth: | Terminabsprache |
| SIEMENScats: | sem_lex(nr,terminabsprache) & nr:rel=terminabsprache&sortal_Terminabsprache(nr) & count_noun_norm(nr) & subst_klasse2_1(nr)terminabsprache & sortal_einigen_auf & count_noun_norm&subst_klasse2_1 |
| SIHUBval: | -- |
| BIgloss: | appointment_scheduling |
| IBMorth: | -- |
| IBMmorph: | -- |
| IBMHUBsyn: | [gender:fem,number:sg,case:ncase_v, syn_ibm: [phon:'Terminabsprache',cuf_macro:common_noun_syn], person:3] |
| TUBsem: | terminabsprache & communicating & - |
| TUEBcomp: | terminabsprache: compound(terminwoche,first(termin), second(absprache), semrel(arg3_rel)). |
| IMSrule: | terminabsprache:[H:terminabsprache(I)] <-> [H:scheduling(I), H1:indef(Y,H2), H2:appointment(Y), H3:of(I,Y)]. |

VERBMOBIL ONLINE LEXICON

Microstructure for collaborative integration of lexica from different research institutes and research departments of companies.

# *Main lexicography Projects since 1989*

Verbmobil

- multilingual lexicographic model for speech-to-speech translation
- Lexicographic infrastructure for German research institutes and companies

ULex:

- Ubiquitous comprehensive online dictionary
- Uyghur case study

Uyo Ibibio Dictionary:

- bilingual dictionary of a Niger-Congo language

ILEX (Integrated Inheritance Lexicon):

- DATR, Inheritance Lexicon Language
  - morphology, prosody, idioms

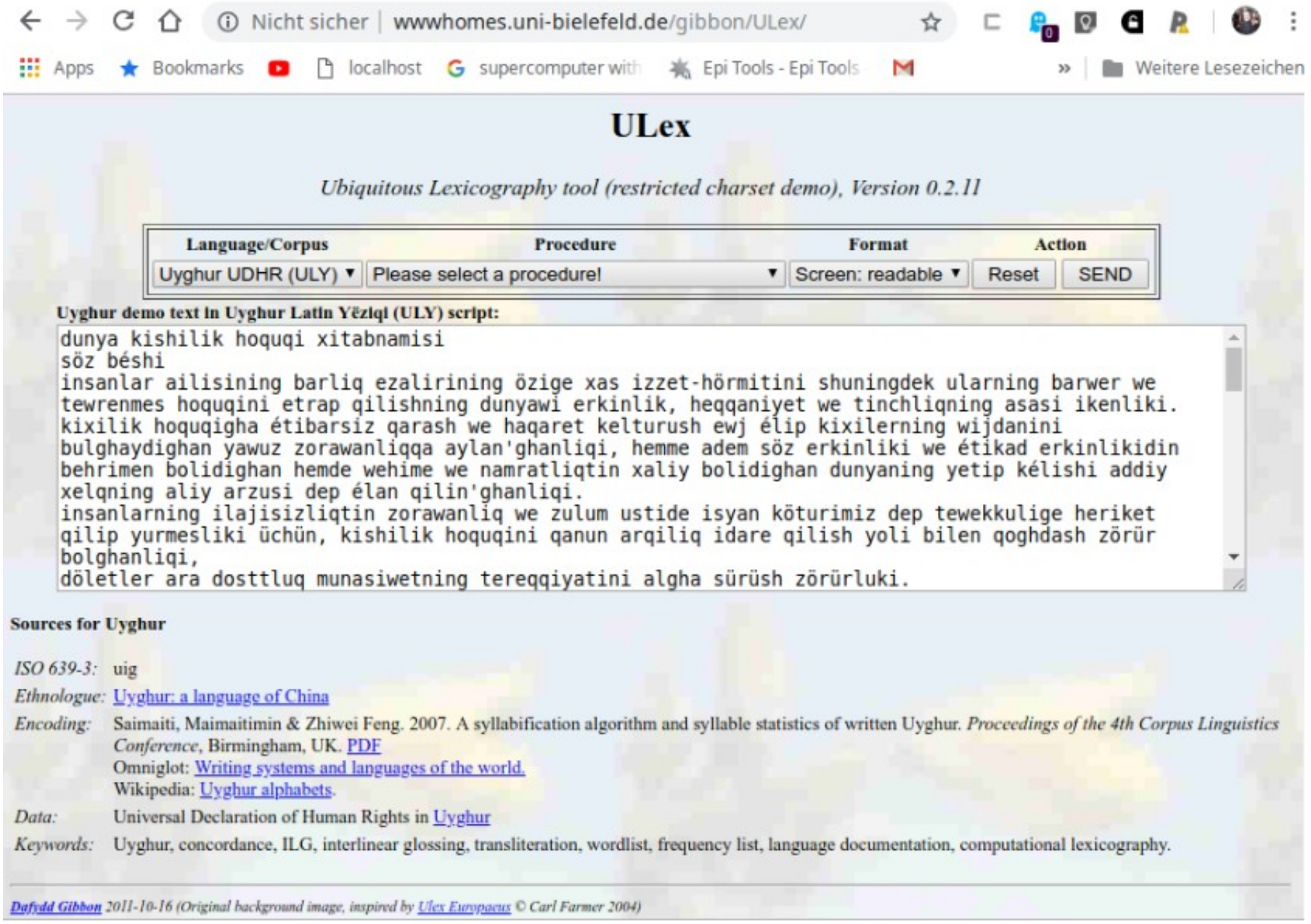
## *Ulex: Ubiquitous Lexicon*

## *Objectives:*

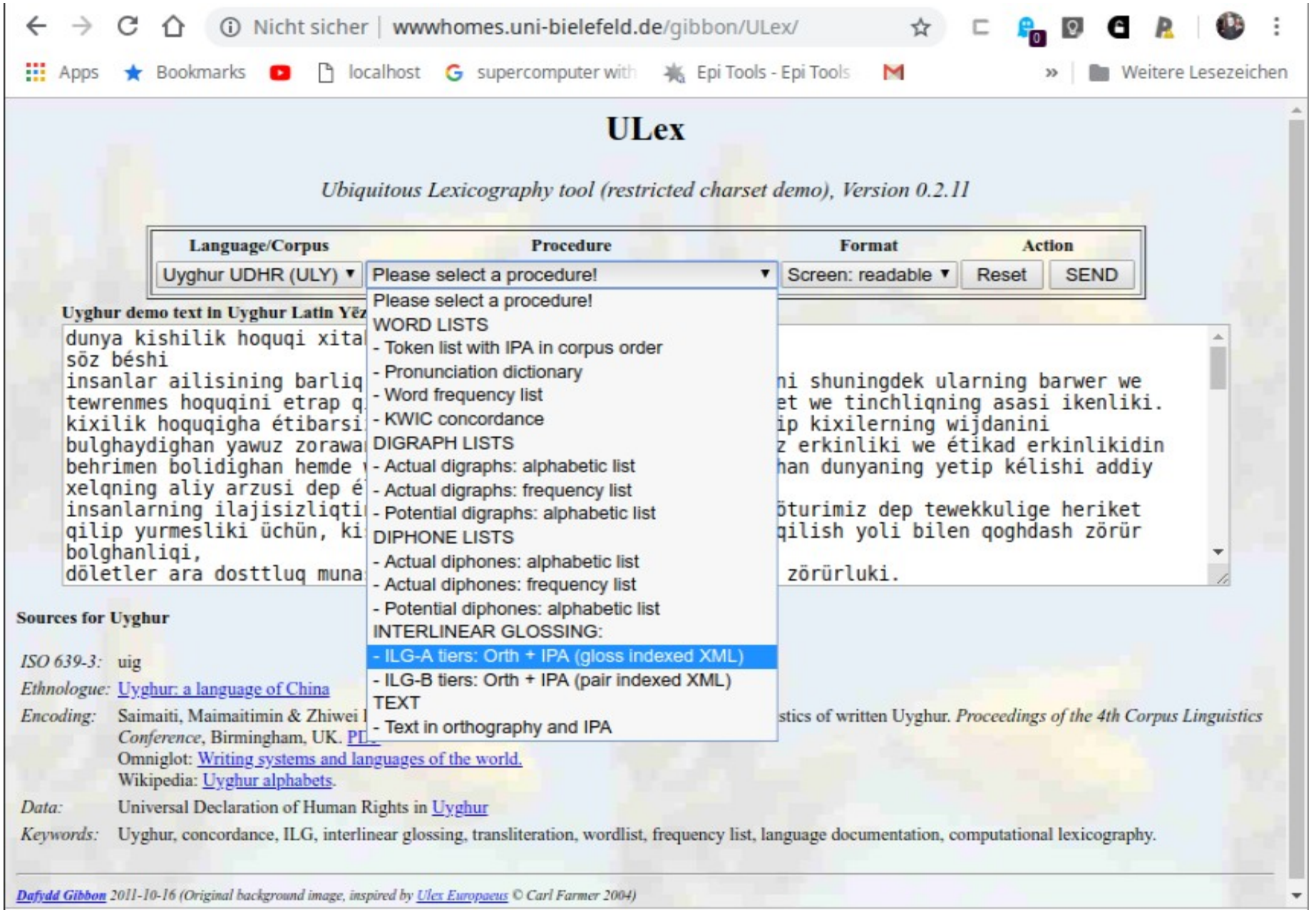
***Language documentation***
***Ubiquitous – Portable***
***Reusable***
***Sustainable***
***Interoperable***

## *Case Study: Uyghur UN declaration*

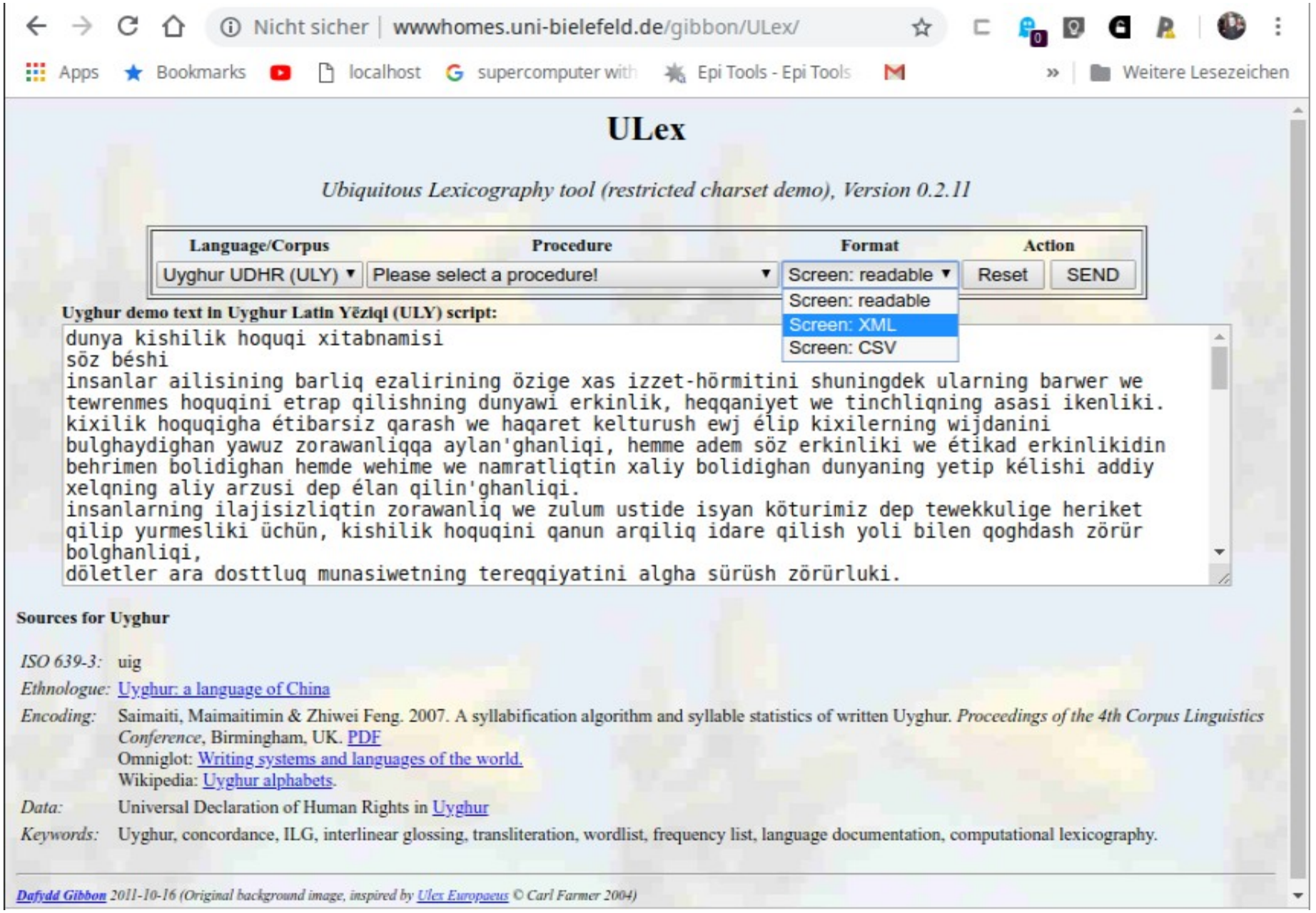
# *Ulex: Ubiquitous Lexicon - Uyghur*



# *Ulex: Ubiquitous Lexicon - Uyghur*



# *Ulex: Ubiquitous Lexicon - Uyghur*



# *Main lexicography Projects since 1989*

Verbmobil

- multilingual lexicographic model for speech-to-speech translation
- Lexicographic infrastructure for German research institutes and companies

ULex:

- Ubiquitous comprehensive online dictionary
- Uyghur case study

Uyo Ibibio Dictionary:

- bilingual dictionary of a Niger-Congo language

ILEX (Integrated Lexicon, Inheritance Lexicon):

- DATR, Inheritance Lexicon Language
  - morphology, prosody, idioms

## *Uyo Ibibio Lexicon*

## *Objectives:*
## *Semasiological dictionary of a local language*
## *Data mining of a legacy print dictionary*
## *Automatic generalisation of online & print dictionaries*

## *Case:*
## *Ibibio, ISO 639-3*

***Niger-Congo > Atlantic-Congo > Volta-Congo > Benue-Congo > Cross River > Delta Cross > Lower Cross***

# *Ibibio – official language of Akwa Ibom State, Nigeria*

Uyo, capital of Akwa Ibom State

# *Ibibio Dictionary*

*ABUILD Language Documentation Curriculum Materials #1*
*COURSE: LEXICOGRAPHY*

**Uyo Ibibio Dictionary**

Eno-Abasi Urua
(Department of Linguistics and Nigerian Languages)
Moses Ekpenyong
(Department of Computer Science)
University of Uyo, Akwa Ibom State, Nigeria
Dafydd Gibbon
(Fakultät für Linguistik und Literaturwissenschaft)
Universität Bielefeld, Germany

***PREPRINT DRAFT V01, 2004-06-13***

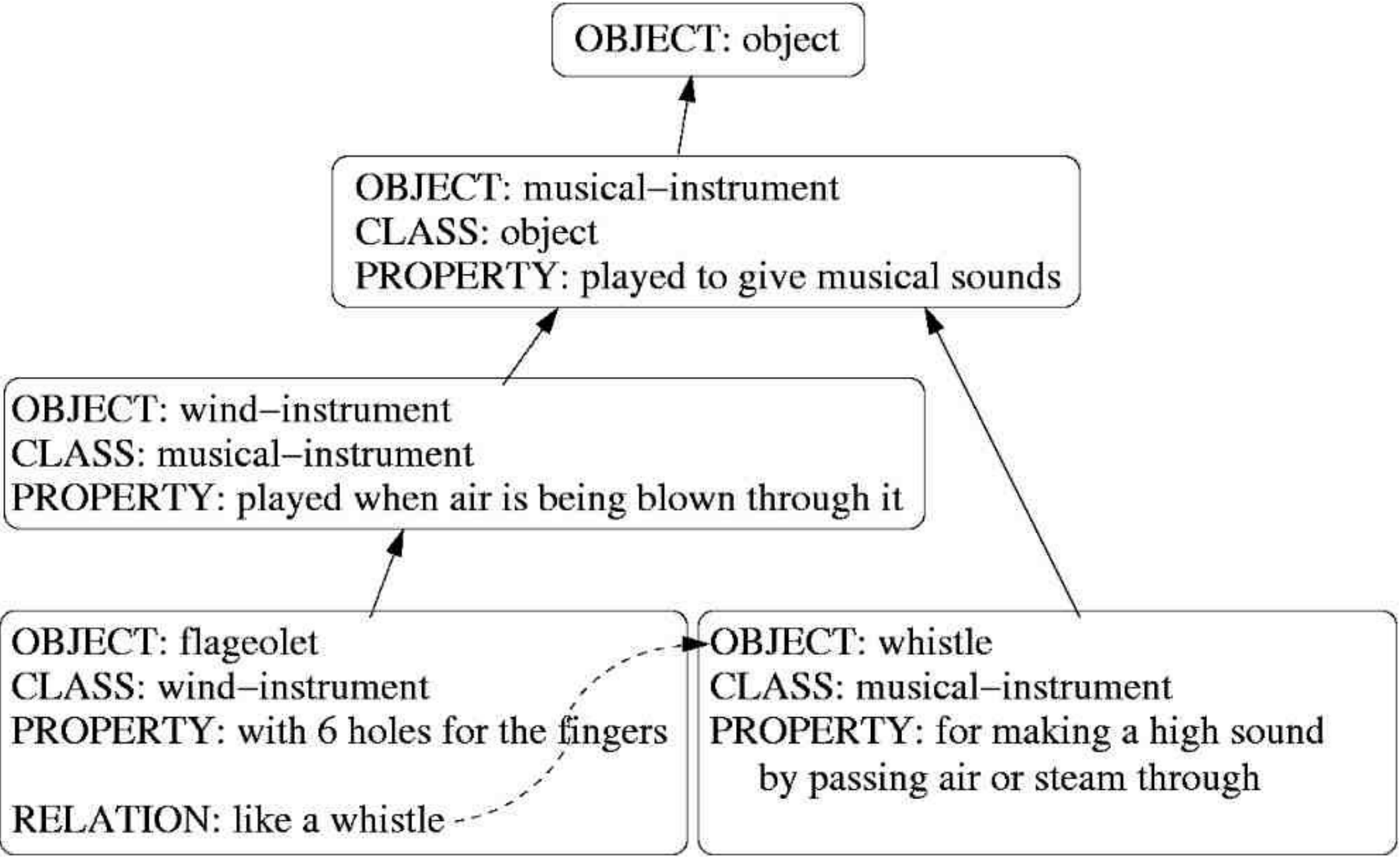
***(created June 14, 2004)***

***ILEX: Integrated Lexicon***

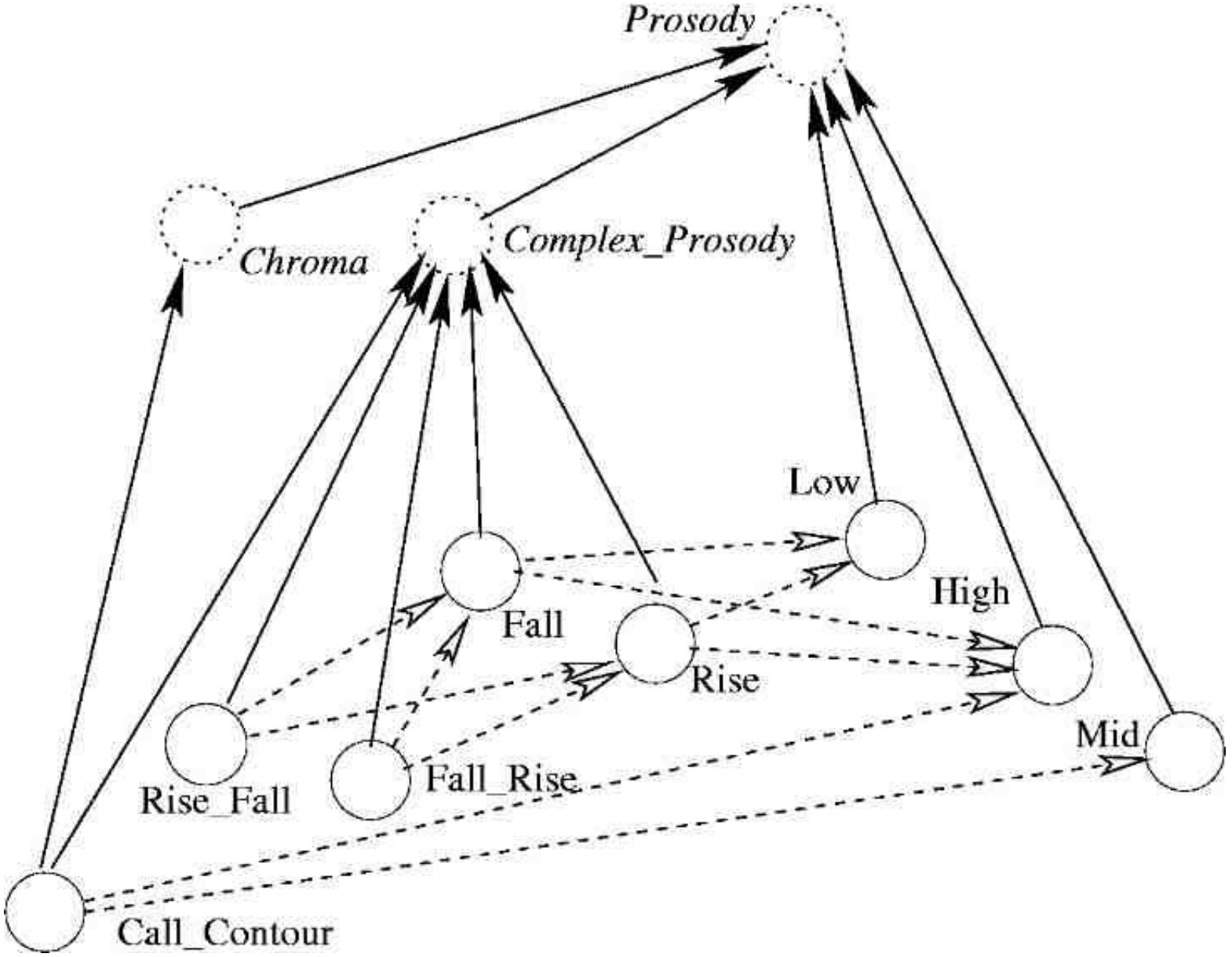
***Objectives:***
***Logical modelling of lexical generalisations***
***Computational integration:***
***macrostructure***
***microstructure***
***mesostructure***

***Cases: idioms, morphology, phonology, prosody***

# *Inheritance Lexicon*



# *ILEX Microstructure for Tonal Inventory: English Call Contour*



# *Lexicon Theory – Lexicon Models – Lexicon Constraints*

# *Lexicon Theory – Lexicon Models – Lexicon Constraints*

Lexicon processing:

- Discovery, acquisition:
  - Induction
  - Abstraction
  - Generalisation
- Access:
  - Technical, ethical, moral, political constraints
  - Database views
  - Semasiological vs. onomasiological views
  - Specialised views:
    - Pronunciation
    - Terminology

Lexicon dissemination

- Standardisation
  - DBMS, CMS
  - ISO
- Commercialisation
- Sustainable format
- Trusted repository
- Interactive vs. Static
- Online vs. Offline
- Print vs. electronic

Gibbon, Dafydd (2005). Spoken language lexicography: an integrative framework. In: Lew Zybatow, ed. (2005), *Translatologie – Neue Ideen und Ansätze*. Frankfurt, etc.: Peter Lang, Europäischer Verlag der Wissenschaften, 247-289.

# *Lexicon Architecture Model*

1. MEGASTRUCTURE

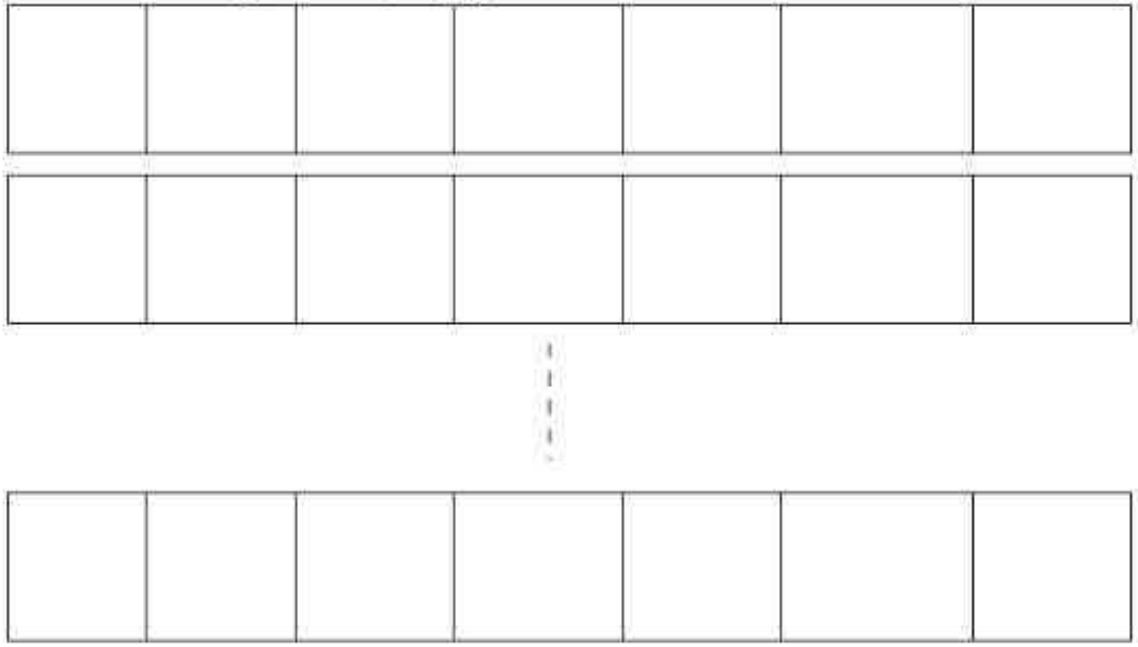
2. METAINFORMATION

Metadata, front matter, back matter

3. MACROSTRUCTURE

4. MICROSTRUCTURE
(data categories, types of lexical information)

Lexical entries

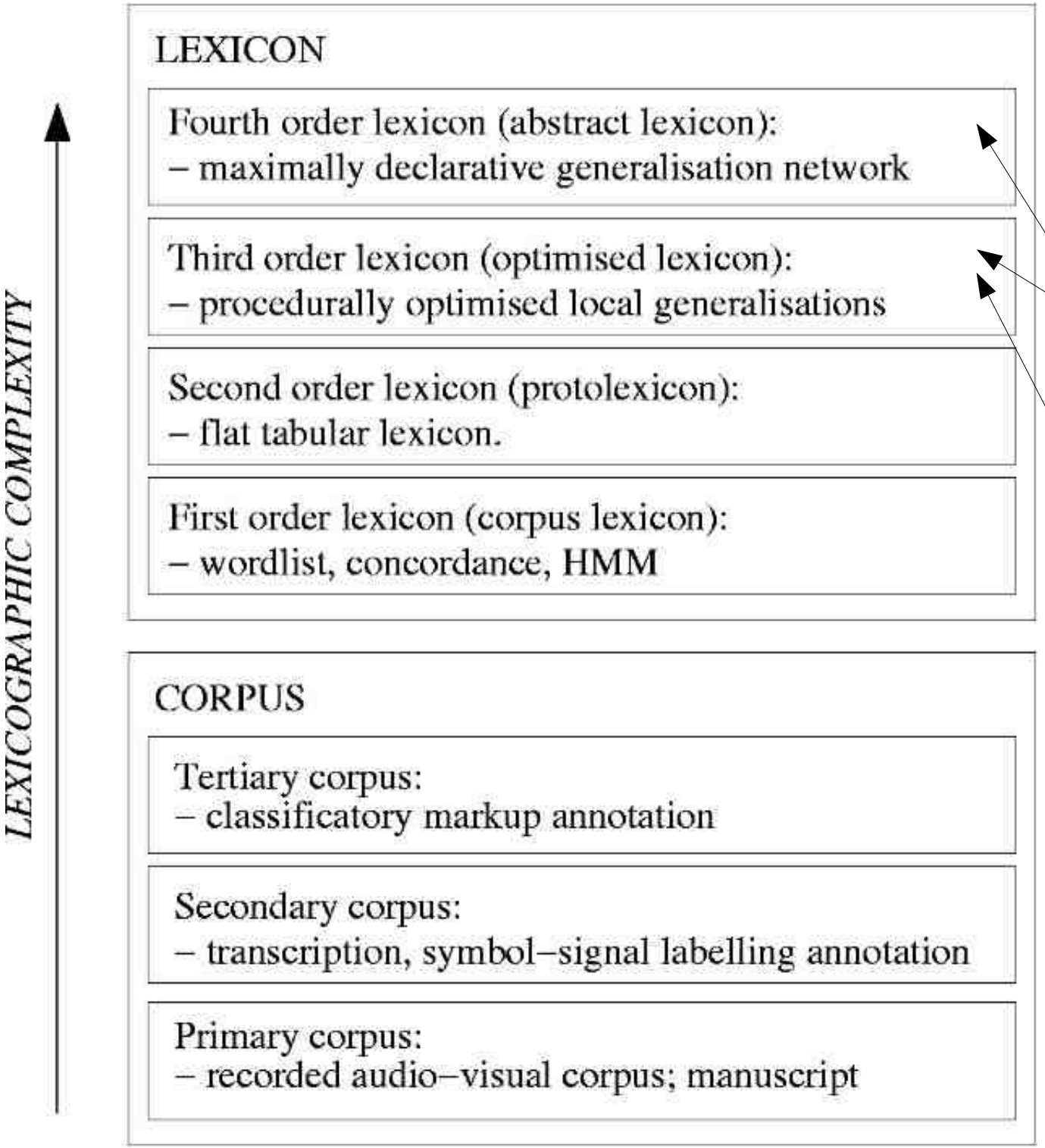


5. MESOSTRUCTURE

Sketch grammar, i.e. conventions for generalisations over microstructure:

- orthography
- pronunciation
- word formation
- syntax
- definitions
- examples

# *Lexicon Discovery Model*




explicit mesostructure:
semasiological thesaurus,
inheritance lexicon

Lemma-structured
standard semasiological
dictionary

# *Summary, Conclusion, Outlook*

- Projects:
  - Projects of different lexicographic sub-genres
  - Use-case orientation
- Principles:
  - Abstraction of shared properties
  - Development of a model of lexicon architecture
  - Development of a model of lexicon acquisition
- Outlook – future essentials:
  - Full-time team effort
  - Intensive coordination
  - Coherent theory basis
  - Efficient computational implementation
  - User satisfaction