

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

# ***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 Lungen (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 Lungen 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 lexicon. *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***



## VM-HyprLex Interface 3

bielefeld.lexdb.v3.3, Mar 18 1996  
(8081 data records, 35 attributes)

String ▼ Terminabsprache KEY type and string

Key ▼ KEY / SubDB SEARCH

Defaults: Consult:lexicon:

Marked ▼ ATTRIBUTE DISPLAY

Coverage Operation

### Morphology, Morphophonology, Morphosemantics

Blorth       Blorthseg       BImorpro       Blorthstem       Blphonstem  
 Biflex       Bilemma       BIsPELL       Blproper       Blcompsem

### Corpus distribution, selection, tagging

BICD1       BICDall       Blpercent       Blrank       Blortherror  
 BLAUBEU       DemoWL       RQH-WL       Blhitlist       FPWL3  
 KIcanon       KIfreq       IMSlem       IMSpos       IMSfreq

### Syntax, Semantics, Transfer, Dialogue, Glossary

SIEMENSorth       SIEMENScats       SIHUBval       Blgloss  
 IBMorth       IBMmorph       IBMHUBsyn  
 TUBsem       TUEBcomp       IMSrule

[Changes](#) - [Reference](#) - [FAQ](#) - [Help doc](#) - [Concordance](#) - [MAIN](#)

VERBMOBIL ONLINE LEXICON

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

<i>Category</i>	<i>Value</i>
<u>BIorth:</u>	Terminabsprache
<u>BIorthseg:</u>	Termin#ab#sprach#+e
<u>BIomorpro:</u>	tE6.m'i:n#?'ap#Spr''a:.x#+@
<u>BIorthstem:</u>	Termin#ab#sprach
<u>BIphonstem:</u>	tE6.m'i:n#?'ap#Spr''a:x
<u>BIflex:</u>	N,akk,sg,fem;N,dat,sg,fem;N,gen,sg,fem;N,nom,sg,fem
<u>BIlemma:</u>	Terminabsprache
<u>BIspell:</u>	--
<u>BIproper:</u>	--
<u>BIcompsem:</u>	ObjEreig
<u>BICD1:</u>	cd1=2_cd12=7_cd3=2_cd4=3_cd5=1
<u>BICDall:</u>	15
<u>BIpercent:</u>	0.00568005%
<u>BIrank:</u>	977
<u>BIortherror:</u>	Termin-Absprache, -
<u>BIAUBEU:</u>	--
<u>DemoWL:</u>	demo-wl
<u>RQH-WL:</u>	--
<u>BIhitlist:</u>	hit#977=15
<u>FPWL3:</u>	fpwl
<u>Kicanon:</u>	tE6m'i:n#Q"ap#Spr"a:x@
<u>Kifreq:</u>	14
<u>IMSlem:</u>	Terminabsprache
<u>IMSpos:</u>	NN
<u>IMSfreq:</u>	8
<u>SIEMENSorth:</u>	Terminabsprache
<u>SIEMENScats:</u>	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
<u>SIHUBval:</u>	--
<u>BIgloss:</u>	appointment_scheduling
<u>IBMorth:</u>	--
<u>IBMorph:</u>	--
<u>IBMHUBsyn:</u>	{gender:fem,number:sg,case:ncase_v, syn_ibm: [phon:'Terminabsprache',cuf_macro:common_noun_syn], person:3}
<u>TUBsem:</u>	terminabsprache & communicating & -
<u>TUEBcomp:</u>	terminabsprache: compound(terminwoche,first(termin), second(absprache), semrel(arg3_rel)).
<u>IMSrule:</u>	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 Lexicography tool (restricted charset demo), Version 0.2.11*

Language/Corpus	Procedure	Format	Action
Uyghur UDHR (ULY) ▾	Please select a procedure! ▾	Screen: readable ▾	Reset SEND

**Uyghur demo text in Uyghur Latin Yëzliqi (ULY) script:**

dunya kishilik hoquqi xitabnamisi  
söz bëshî  
insanlar ailisining barliq ezalirining özige xas izzet-hörmitini shuningdek ularning barwer we  
tewrenmes hoquqini etrap qilishning dunyawî erkinlik, heqqaniyet we tinchliqning asasi ikenliki.  
kixilik hoquqigha étibarsiz qarash we haqaret kelturush ewj élip kixilerning wijdanini  
bulghaydighan yawuz zorawanliqqa aylan'ghanliqi, hemme adem söz erkinliki we étikad erkinlikidin  
behrimen bolidighan hemde wehime we namratliqtin xaliy bolidighan dunyaning yetip kélishi addiy  
xelqning aliy arzusi dep élan qilin'ghanliqi.  
insanlarning ilajisizliqtin zorawanliq we zulum ustide isyan köturimiz dep tewekkulige heriket  
qilip yurmesliki üçhün, kishilik hoquqini qanun arqiliq idare qilish yoli bilen qoghdash zörür  
bolghanliqi,  
döletler ara dosttluq munasiwetning tereqqiyatini algha sürüş zörürluki.

**Sources for Uyghur**

ISO 639-3: [uig](#)

Ethnologue: [Uyghur: a language of China](#)

Encoding: Saimaiti, Maimaitimin & Zhiwei Feng. 2007. A syllabification algorithm and syllable statistics of written Uyghur. *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, UK. [PDF](#)

Omniglot: [Writing systems and languages of the world](#).

Wikipedia: [Uyghur alphabets](#).

Data: Universal Declaration of Human Rights in [Uyghur](#)

Keywords: Uyghur, concordance, ILG, interlinear glossing, transliteration, wordlist, frequency list, language documentation, computational lexicography.

[Dafydd Gibbon](#) 2011-10-16 (Original background image, inspired by [Ulex Eurypactus](#) © Carl Farmer 2004)

# Ulex: Ubiquitous Lexicon - Uyghur

The screenshot shows the Ulex web application interface. At the top, the browser address bar displays "wwwhomes.uni-bielefeld.de/gibbon/ULex/". The page title is "Ulex" and the subtitle is "Ubiquitous Lexicography tool (restricted charset demo), Version 0.2.11".

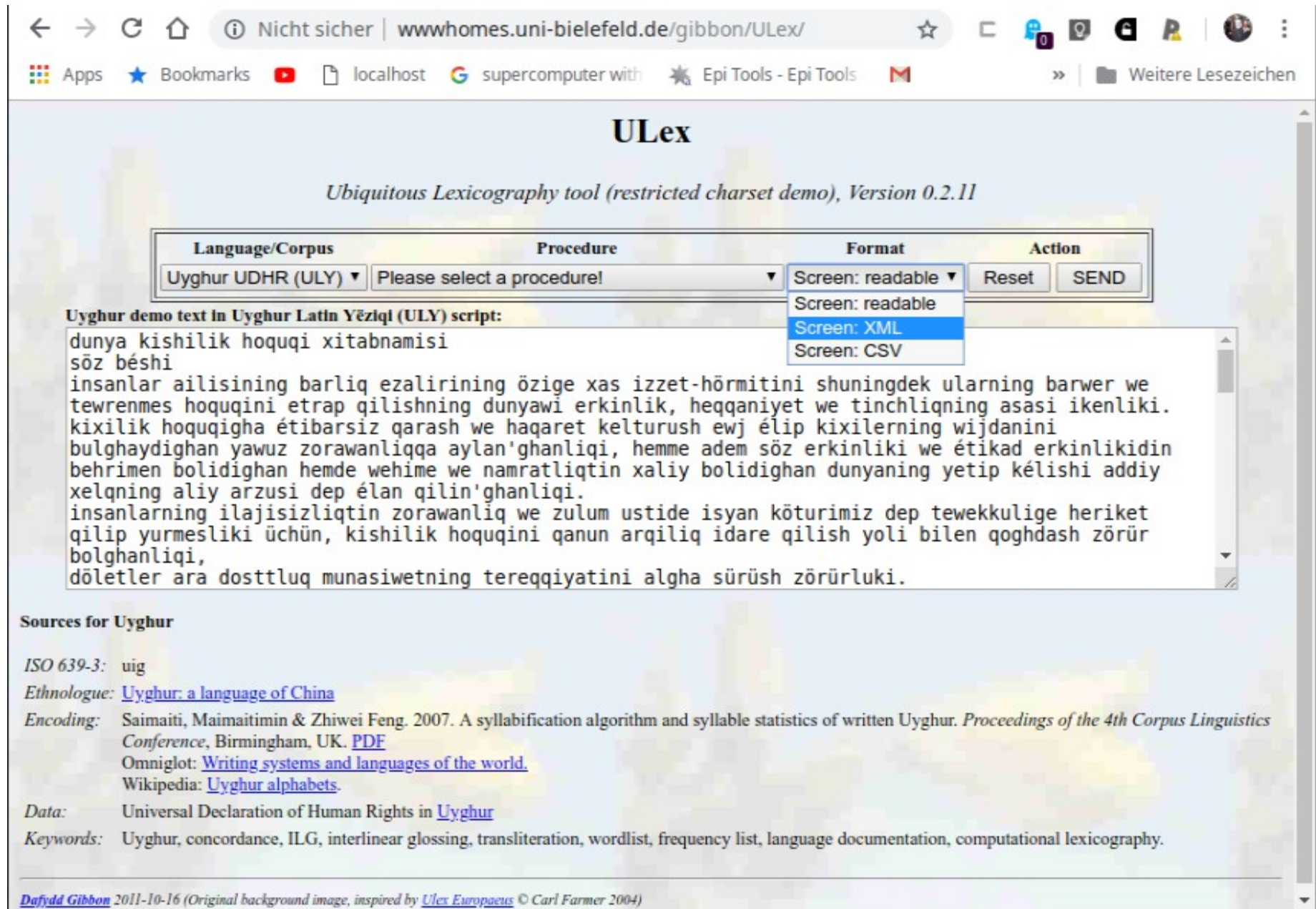
The main interface consists of a form with four columns: "Language/Corpus", "Procedure", "Format", and "Action".

- Language/Corpus:** A dropdown menu is open, showing "Uyghur UDHR (ULY)". Below it, a text area contains the text: "Uyghur demo text in Uyghur Latin Yëz dunya kishilik hoquqi xita söz béshi insanlar ailisining barliq tewrenmes hoquqini etrap q kixilik hoquqigha étibarsi bulghaydighan yawuz zorawa behrimen bolidighan hemde xelqning aliy arzusi dep é insanlarning ilajisizliqti qilip yurmesliki üçün, ki bolghanliqi, döletler ara dosttluq muna".
- Procedure:** A dropdown menu is open, showing "Please select a procedure!". The menu items are: "WORD LISTS", "DIGRAPH LISTS", "DIPHONE LISTS", and "INTERLINEAR GLOSSING:". The "INTERLINEAR GLOSSING:" section is expanded, showing options: "- ILG-A tiers: Orth + IPA (gloss indexed XML)", "- ILG-B tiers: Orth + IPA (pair indexed XML)", "TEXT", and "- Text in orthography and IPA".
- Format:** A dropdown menu is open, showing "Screen: readable".
- Action:** Two buttons: "Reset" and "SEND".

Below the form, there is a section titled "Sources for Uyghur" with links to ISO 639-3, Ethnologue, Encoding, Omniglot, and Wikipedia. At the bottom, there is a "Data" section and a "Keywords" section.

At the very bottom, a footer note reads: "Dafydd Gibbon 2011-10-16 (Original background image, inspired by Ulex Europaetus © Carl Farmer 2004)".

# Ulex: Ubiquitous Lexicon - Uyghur



The screenshot shows the Ulex web application interface. At the top, the browser address bar displays "Nicht sicher | wwwhomes.uni-bielefeld.de/gibbon/ULex/". The page title is "Ulex" and the subtitle is "Ubiquitous Lexicography tool (restricted charset demo), Version 0.2.11".

The main interface consists of a form with four columns: "Language/Corpus", "Procedure", "Format", and "Action". The "Language/Corpus" dropdown is set to "Uyghur UDHR (ULY)". The "Procedure" dropdown is set to "Please select a procedure!". The "Format" dropdown is set to "Screen: readable", and a dropdown menu is open showing options: "Screen: readable", "Screen: XML", and "Screen: CSV". The "Action" column contains "Reset" and "SEND" buttons.

Below the form, the text area displays the following Uyghur text in Latin script:

Uyghur demo text in Uyghur Latin Yēzliqi (ULY) script:  
dunya kishilik hoquqi xitabnamisi  
söz béshi  
insanlar ailisining barliq ezalirining özige xas izzet-hörmitini shuningdek ularning barwer we  
tewrenmes hoquqini etrap qilishning dunyawî erkinlik, heqqaniyet we tinchliqning asasi ikenliki.  
kixilik hoquqigha étibarsiz qarash we haqaret kelturush ewj élip kixilerning wijdanini  
bulghaydighan yawuz zorawanliqqa aylan'ghanliqi, hemme adem söz erkinliki we étikad erkinlikidin  
behrimen bolidighan hemde wehime we namratliqtin xaliy bolidighan dunyaning yetip kélishi addiy  
xelqning aliy arzusi dep élan qilin'ghanliqi.  
insanlarning ilajisizliqtin zorawanliq we zulum ustide isyan köturimiz dep tewekkulige heriket  
qilip yurmesliki üçün, kishilik hoquqini qanun arqiliq idare qilish yoli bilen qoghdash zörür  
bolghanliqi,  
döletler ara dosttluq munasiwetning tereqqiyatini algha sürüş zörürluki.

Below the text area, there is a section titled "Sources for Uyghur" with the following information:

ISO 639-3: uig  
Ethnologue: [Uyghur: a language of China](#)  
Encoding: Saimaiti, Maimaitimin & Zhiwei Feng. 2007. A syllabification algorithm and syllable statistics of written Uyghur. *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, UK. [PDF](#)  
Omniglot: [Writing systems and languages of the world](#).  
Wikipedia: [Uyghur alphabets](#).  
Data: Universal Declaration of Human Rights in [Uyghur](#)  
Keywords: Uyghur, concordance, ILG, interlinear glossing, transliteration, wordlist, frequency list, language documentation, computational lexicography.

At the bottom, there is a footer: [Dafydd Gibbon](#) 2011-10-16 (Original background image, inspired by [Ulex Europaetus](#) © Carl Farmer 2004)

# ***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:***

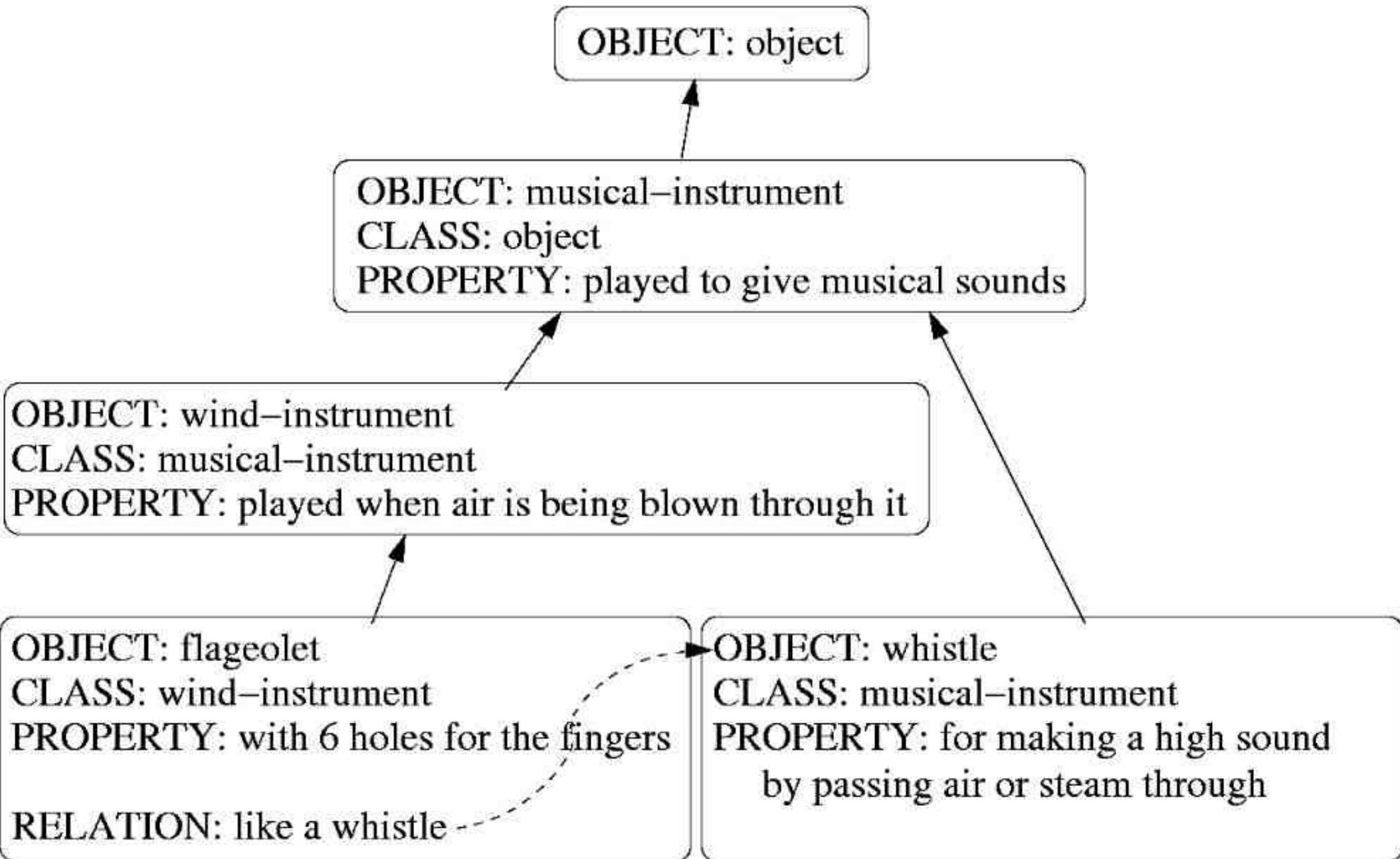
***macrostructure***

***microstructure***

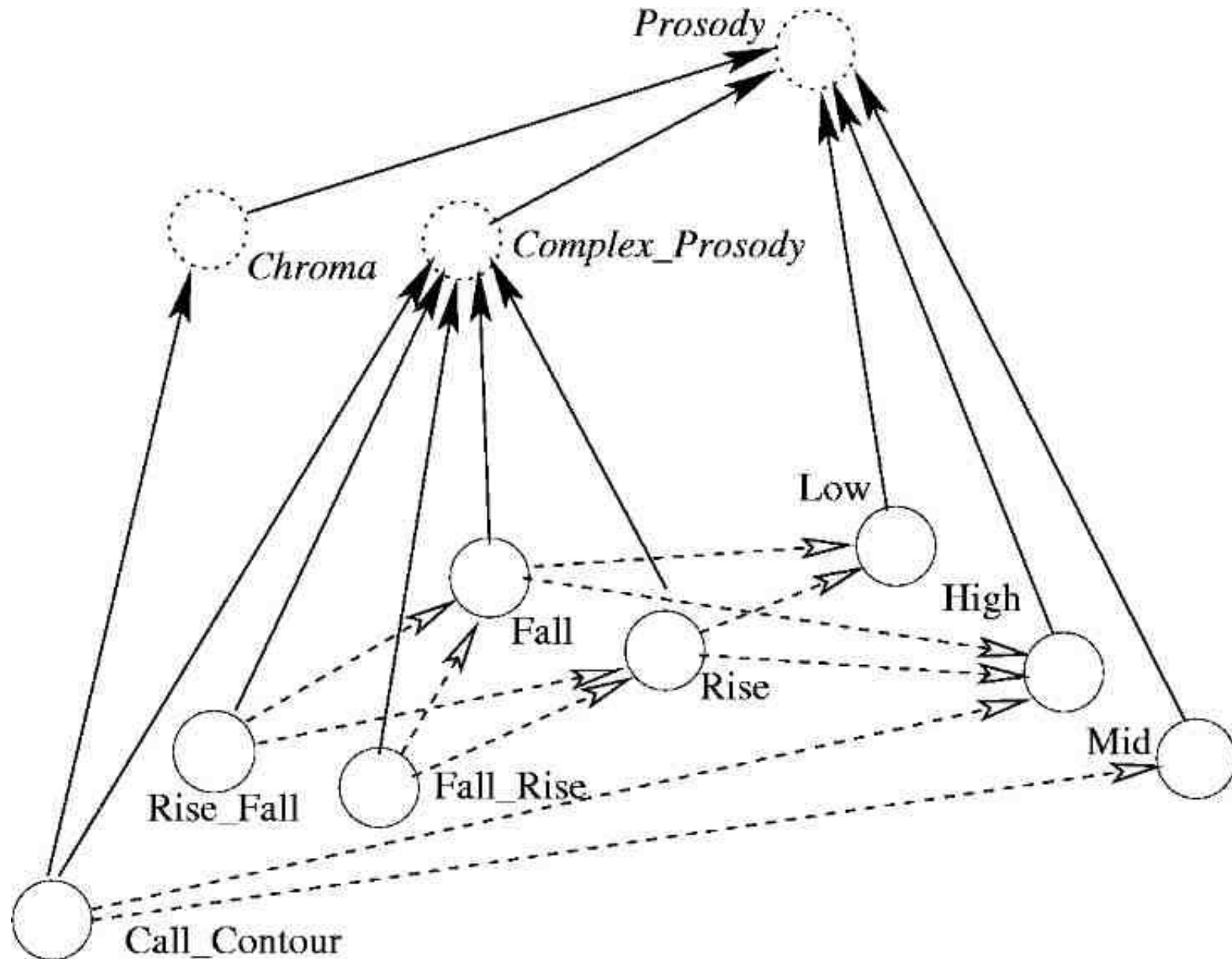
***mesostructure***

***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

Fourth order lexicon (abstract lexicon):  
 – maximally declarative generalisation network

Third order lexicon (optimised lexicon):  
 – procedurally optimised local generalisations

Second order lexicon (protollexicon):  
 – flat tabular lexicon.

First order lexicon (corpus lexicon):  
 – wordlist, concordance, HMM

CORPUS

Tertiary corpus:  
 – classificatory markup annotation

Secondary corpus:  
 – transcription, symbol–signal labelling annotation

Primary corpus:  
 – recorded audio–visual corpus; manuscript

**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