The Bielefeld–Abidjan documentation project:

Information Types and Dissemination Media

Dafydd Gibbon

LSA Linguistic Exploration Workshop

6 January 2000

with particular thanks to

François Adopo, Firmin Ahoua, Jérémie Kouadio
Christian Lehmann, Ulrich Dausendschön–Gay
Sandrine Adouakou, Guy Kaul

for their close cooperation

Abstract

The Bielefeld-Abidjan documentation project 'Encyclopédie des Langues de Côte d'Ivoire' is funded by the Deutscher Akademischer Austauschdienst (DAAD, German Academic Exchange Service) with the goal of developing a specification and design for a multimedia encyclopedia of the four language groups of the Ivory Coast (Kru, Kwa, Mande, Gur). Initially, the Kwa languages are being focussed on, in particular the closely related langauges Baule and Anyi. Four exploration methods are being developed:

1. Detailed questionnaire on structural and functional aspects of language, with CGI-based form input and hypertext output.
2. Systematic acoustic phonetic analyses to extend the phonological analyses under (1).
3. Specialised phonetic databases, e.g. diphone database for testing models via speech synthesis.
4. Text collection with
   - audio
   - transcription
   - annotation (selected)
   - illustrated lexicon
   - HyprLex dynamic concordance (for models, see "http://coral.lili.uni-bielefeld.de/HyprLex/")

The prototype tools for these tasks are being made available for use in the field. Close attention has been paid to appropriate levels of analysis for tools in this area. For example, the variety of mutually incompatible fonts for African languages, IPA, etc., involve much unnecessary engineering overhead in automatical processing, e.g. in concordance generation, and are not conducive to free information interchange. General computational principles suggest that this is an overly concrete level of representation and that a more generic and standardisable platform–neutral transcription format, such as the X–SAMPA machine–readable (ASCII) IPA coding, is more suitable, with conversion to reader–friendly fonts at a post–processor stage. We have developed X–SAMPA encodings for Bete (Kru) and Abbey (Kwa), and by the end of the project will have codings for the other languages currently being treated.

The phonetician's creed

The specific goals and modes of expression change as society changes, but ...

661. I take it as axiomatic that every man is endowed with certain abilities which he can use for the general good. People with ability for applied phonetics have in a particularly high degree the capacity for doing "useful work" — by which term I mean work "conducive to the ultimate well–being of humanity". In contrast with workers in some other scientific fields, they can keep in view throughout their investigations a definite humanistic object towards which they can direct their attention, namely, the improvement of the means of oral and written communication between man and man. The findings of phonetic science give people skill in communicating effectively with each other.

Daniel Jones (1950): *The Phoneme*, Cambridge: Heffer, Pp. 217f.

Goals, particpants, institutions

Title: 'Encyclopédie des Langues de Côte d'Ivoire'

Goal: Development of specification, design and prototype implementation demonstration of a multimedia encyclopedia of the four language groups of the Ivory Coast (Kru, Kwa, Mande, Gur), initially Kwa (Baule and Anyi).

Funded by: Deutscher Akademischer Austauschdienst (DAAD, German Academic Exchange Service)

Coordinating partner: Universität Bielefeld, Germany (Dafydd Gibbon, Ulrich Dausendschön–Gay, Martina Drescher, Rolf Ehnert), in cooperation with Universität Erfurt (Christian Lehmann)

Local partners: Institut de Linguistique Appliquée and Département de Linguistique, Université de Cocody, Abidjan, Côte d'Ivoire (esp. François Adopo, Jérémie Kouadio, Eddy Gbéry, François Kipré Blé)

Querying some assumptions – 1

A cherished tenet of traditional 'Western Science':

Achievement is *individual*, measurable on a scale from *genius* to *dunce* ...

- *genius*, *hero*, *saint*, ...
  - recognised, e.g. Nobel (or other) Prize winner ...
  - belatedly recognised, e.g. Boole, Frege, ...
  - unrecognised, i.e. one's favourite model, such as Henry Sweet, Roger Kingdon, ...
- *first rate* scientist, artist, ...
- *second rate* ...
- *third rate* ...
- ...
- ...
- *dunce*

Querying some assumptions – 2

Thesis: The *genius-dunce model* is an entirely inappropriate model for the global information society of the 20th century, because ...

1. science is a social paradigm with a variety of local conventions and patronages;
2. concentration on individual achievement and failure exposes the community to *divide et impera* strategies which support political manipulation;
3. in many societies – perhaps most – collective achievement is valued more highly than invidividual achievement;
4. in the contemporary global information society individual achievement without collective achievement is impossible.

Corollary 1: Importation of individual focus can have highly negative consequences in a collectively focussed society.

→ provokes local sanctions, political particularism and elitism.

Corollary 2: Individual achievement is a necessary but not a sufficient component of collective achievement.

→ Cooperative consortial research and team training are the appropriate methods for the 21st century.

Tendency: Global SOCIAL (politico-economic, educational) and TECHNOLOGICAL (informational – GSM, Internet) structures are steadily undermining the *genius-dunce model*.

Another tenet: 'Linguistic science OR linguistic engineering'?

A basic insight:

> The joy of scientific innovation in detail
> should not blind one to the fact that
> not only is technology dependent on
> scientific achievement
> but, increasingly,
> scientific achievement is dependent on
> technological achievement

An informal but necessary working assumption:

> Modern linguistic fieldwork
> contains a large
> *team engineering component*.

A necessity: 'Linguistic science AND linguistic engineering'?

Inseparable methodologies ...

1. Use of engineering infrastructure:
   software, hardware, periphery, network
2. Applications-oriented:
   (a) Scientific research is itself an application of technology.
   (b) The language community itself should be adequately repaid.
3. Engineering criteria:
   (a) More than innovative good invidual ideas: innovative couplings of proven results.
   (b) More than proof of formal soundness and completeness: empirical soundness and completeness.
   (c) More than proof of functionality: scalability, robustness, usability by third parties.
4. Application of goal-oriented software engineering procedures, such as
   (a) Requirements specification document (tasks, contents, inputs, outputs)
   (b) Design study (architecture, interfaces, algorithms, data structures)
   (c) Implementation (choice of platform, programming languages, formats)
   (d) Evaluation (black box; glass box)
5. Large scale consortial project organisation (over and above individual research).

Requirements specification

Deliverable:

Encyclopedia: Specimen encyclopedia entry / entries for selected language(s).

Formats: Standard transcription, annotation, database, questionnaire, text formats.

Platform: Inexpensive, *de facto* standard (e.g. PC, MS, browser)

Dissemination: Paper, CD-ROM, web.

Maintenance: Future project(s) to be applied for.

User group: Local and global scientific communities; local education planners.

Development group: Desc., comp., app. linguists.

Tasks: Collation of different data types:

Typology questionnaire: Structural and functional properties, situation of language (Lehmann)

Corpus database: Illustration of phonological / tonological system, morphological and grammatical constructions; oral interaction, incl. oral literature; specialised data for experimental purposes (laryngograph; diphone); transcriptions and annotations.

Lexical database: HyprLex model for microstructure with transcription, concordance links, graphics, audio.

Time line: Four–year prototype development limit on limited funding (travel and student exchange only):

Year 1: Joint development of questionnaire, initial applications.

Year 2: Training in the field, and student exchange for test questionnaire completion.

Year 3: Training in the field, and student exchange for phonetic database construction.

Year 4: Implementation of encyclopaedia study.

Design criteria

Deliverable: hyperdocument

Structure – components, data types:

- Typological questionnaire
- Acoustic and text corpus database
- Hyperlexicon (in final version: SGML/XML)
- Standards specifications (transcriptions, orthography, formats, ...)

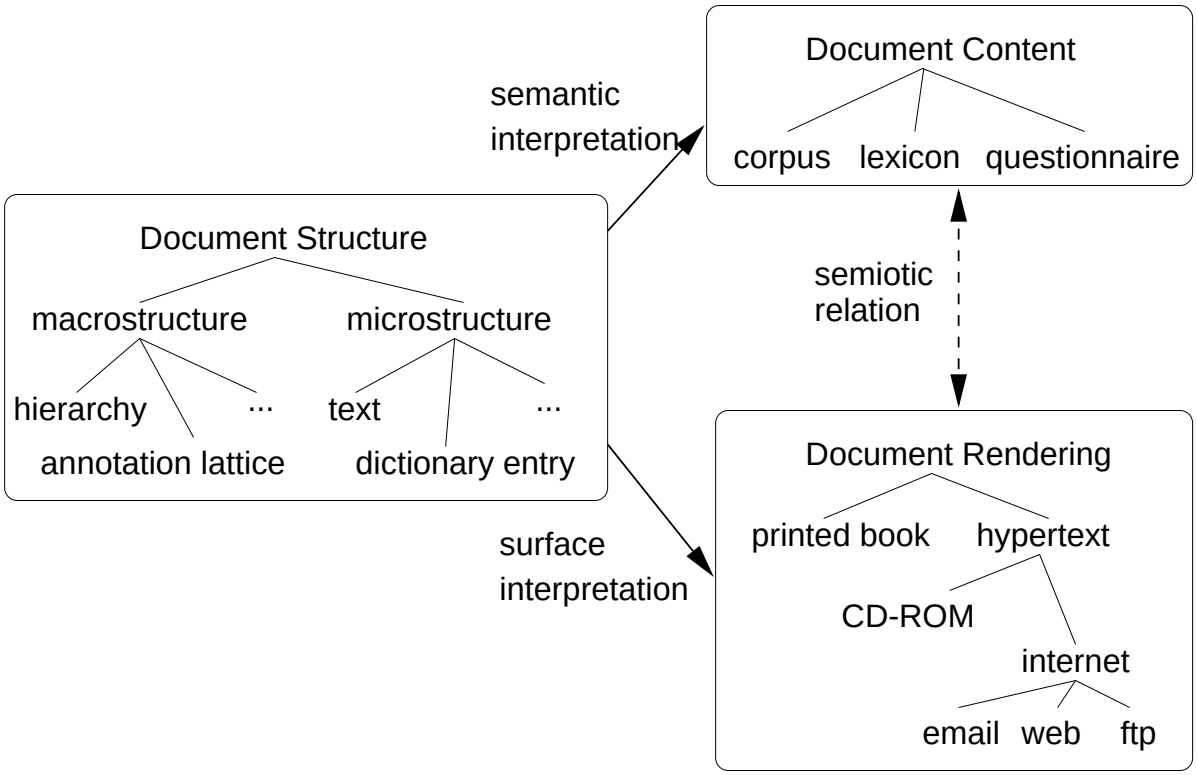
Interaction – access, navigation:

- Server-client architecture (implementation: CGI, possibly mSQL)
- Map/menu-oriented navigation (implementation: JavaScript/HTML)
- Form-oriented data input (implementation: JavaScript/HTML)
- Menu-controlled on-demand generation of hypertext and paper output
- Experimental applications: speech synthesis

Logistics: See task list and time line.

Design: A text linguistic model

Declarative dimension:



Procedural / operational dimension:

- Acquisition of Document Structure from Document Content
- Generation of Document Rendering from Document Structure
- Encoding of document structure in a higher–level language
  - relational database (currently UNIX ASCII)
  - linguistic AVM formalism
  - linguistic inheritance formalism (DATR theory induced from relDB)
  - markup language (e.g. XML)

Design: Document Content, Data Sorts

Corpus:

1. Technical data sorts:
   (a) Recording
   (b) Transcription
   (c) Alignment
   (d) Annotation
2. Linguistic data sorts:
   (a) Oral literature; riddles
   (b) Everyday interaction; games
   (c) Interview
   (d) Linguistic
   (e) Specialised

Lexicon:

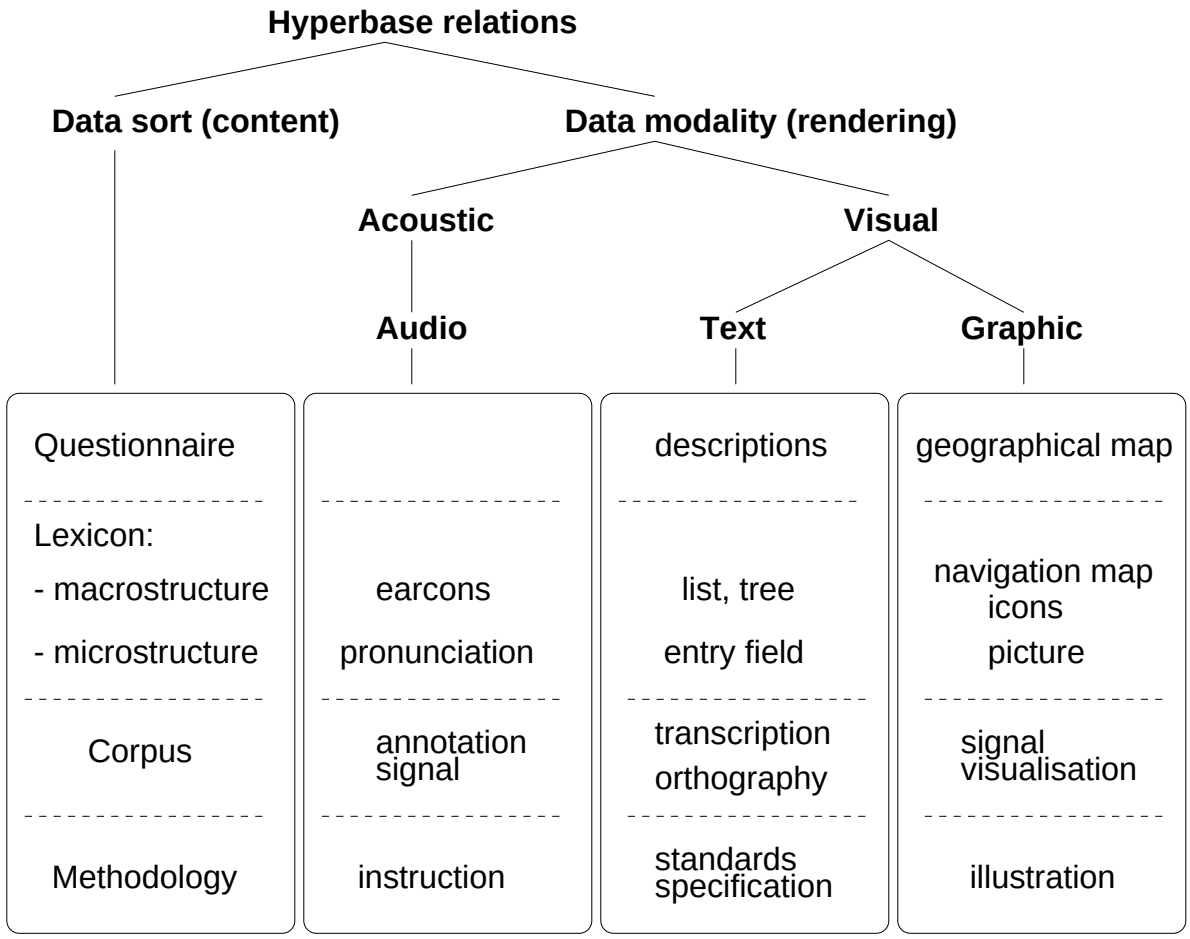
1. Microstructural data sorts:
   (a) Orthography
   (b) Phonology ($\rightarrow$ database)
   (c) Tonology ($\rightarrow$ database)
   (d) Morphology ($\rightarrow$ database)
   (e) Syntax ($\rightarrow$ database)
   (f) Context ($\rightarrow$ concordance)
   (g) Meaning ($\rightarrow$ domain taxonomy)
   (h) Gloss
   (i) Language
   (j) Maintenance
2. Macrostructural data sorts:
   (a) Concordance (-¿ database)
   (b) Domain taxonomy

Questionnaire:

1. Free form data
2. Rule format content
3. Standard generic reference (e.g. IPA)

Design: Data Sorts & Data Modalities

Direct semiotic relations between data sorts and data modalities, bypassing the document structure level, can be represented heuristically as a table:



Implementation: coding problems

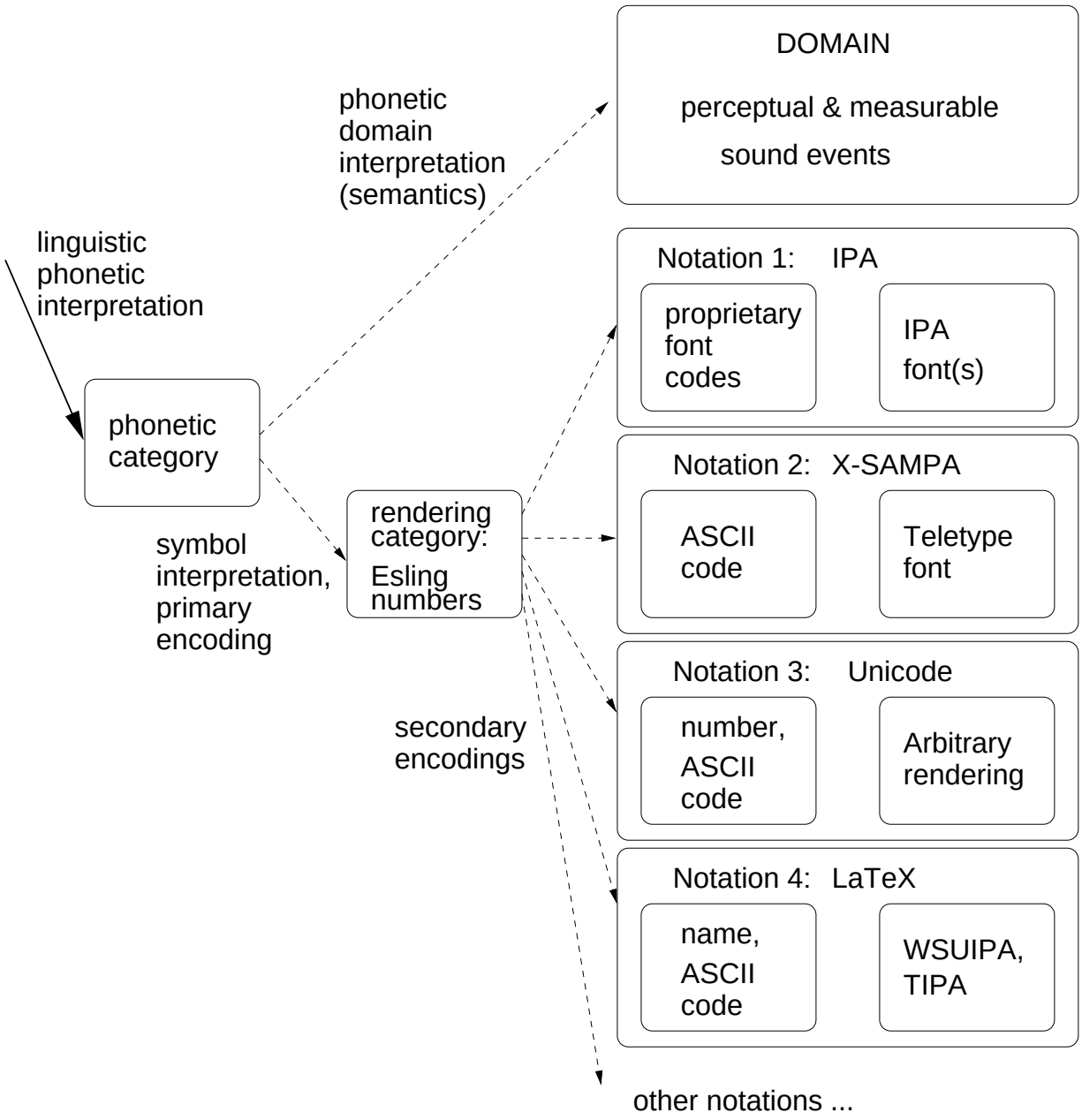
- Variety of individualistic orthographies developed by teachers, priests and linguistics for local language varieties.
- Tendency to phonemic rather than more abstract morphophonemic encoding.
- Attachment to particular symbol shapes.
- Choice of IPA symbol shapes leads to
  - unnecessary engineering overhead in automatic processing;
  - incompatibilities due to proprietary fonts.

A practical problem:

It is not always easy to persuade co-workers to abstract from specific symbol shapes.

Implementation: generic coding

Many codings used in spoken language engineering are language specific, or even lab specific. For efficient, portable documentation, a generic approach to coding is required:



Implementation: practical coding

Encoding criteria include

- 'Any nambiguous manner of representing pronunciation by means of writing is called a "phonetic transcription" '. (Daniel Jones, 1950)
- Typographical broadness: phonemic (broad phonetic) transcriptions can be typographically broad – and avoid inconvenient characters.
- Human readability: encodings need to be understood and reproduced by human encoders and decoders.
- Machine readability: encodings need to be portable, non-proprietary, easily processed, for example with common ASCII oriented scripting languages.

The ASCII based encodings fulfil these criteria. Currently X-SAMPA is the most versatile in terms of

- human readability,
- machine readability,
- ease of translation into other ASCII–represented secondary encodings such as Unicode, LaTeX.

We have developed X-SAMPA encodings for a number of Ivorian languages, and are extending these.

A caveat: transcription is not orthography:

- Typographical broadness: it would be convenient for the language documenter if orthography, whether phonemic or morphophonemic, were also typographically broad.
- Typographical narrowness: But it isn't: orthography contributes to group identities and individual features may be highly valued in encoding phonological categories.

X–SAMPA coding for Adiukru (Kwa)

C	labial	alveolar	palatal	velar	labiovelar
[-cont,-voice]	p	t	tS	k	k_p
[-cont,+voice]	b	d	dZ	g	g_b
[+cont]	f	s	j	h	w
[+nas]	m	n	J	N	
[+lat]		l			
[-trill]		r			

V	Anterior	Posterior
Close	i	u
Half-close	e	o
Half-open	E	O
Open	a	

T	
Simple	Complex
L ‘	HL ’‘
H ’	LH ‘’

Note:

Tone marking with accents is not standard X-SAMPA.

X–SAMPA coding for Anyi (Kwa)

Note that the Indienié dialect of Anyi represented has no labiovelars; allophone pairs are indicated by the slash separator.

C	labial	alveolar	palatal	velar/glottal
[-cont,-voice]	p	t	c	k
[-cont,+voice]	b	d	j\	g
[+cont]	f/v	s/z		h
[+lat]		l/r		
[+nas]	m	n	J	N
[-C -V]	H		j	w

V	Anterior +ATR	Anterior -ATR	Central	Posterior +ATR	Posterior -ATR
Close -nas	i	I		u	U
Close +nas	ĩ	Ĩ		ũ	Ũ
Mid	e	E		o	O
Open -nas			a		
Open +nas			ã		

T	
Simple	Complex
L ‘	HL ’‘
H ’	LH ‘’

Note:

Tone marking with accents is not standard X-SAMPA.

Implementation: Corpus

Corpus containing

- audio (standard format)
- transcription (modified X-SAMPA)
- annotation (standard format)
- LexDB with VerbMobil HyprLex tools
- HyprLex dynamic concordance:
  `"http://coral.lili.uni-bielefeld.de/HyprLex/"`
- Specialised phonetic databases, e.g.
  - diphone database for testing models via speech synthesis,
  - test suites for perceptual and other experiments.

Implementation: Anyi database draft

Specification:

- Domain: 'Corps humain'.
- Field selection:
  1. Transcription: modified X-SAMPA (tones marked with accents)
  2. Skeleton: Consonants, Nasals, Liquids, Vowels
  3. Tone: H, L, M
  4. Gloss: French
  5. Audio file: .wav format, currently 22050 Hz

  Other fields contain: further linguistic information types, domain graphics (illustrations) and phonetic graphics (pitch trace, waveform, spectrogram), concordance reference, etc.
- Automatic extraction:
  Skeleton and Tone fields are extracted automatically from the X-SAMPA transcription.

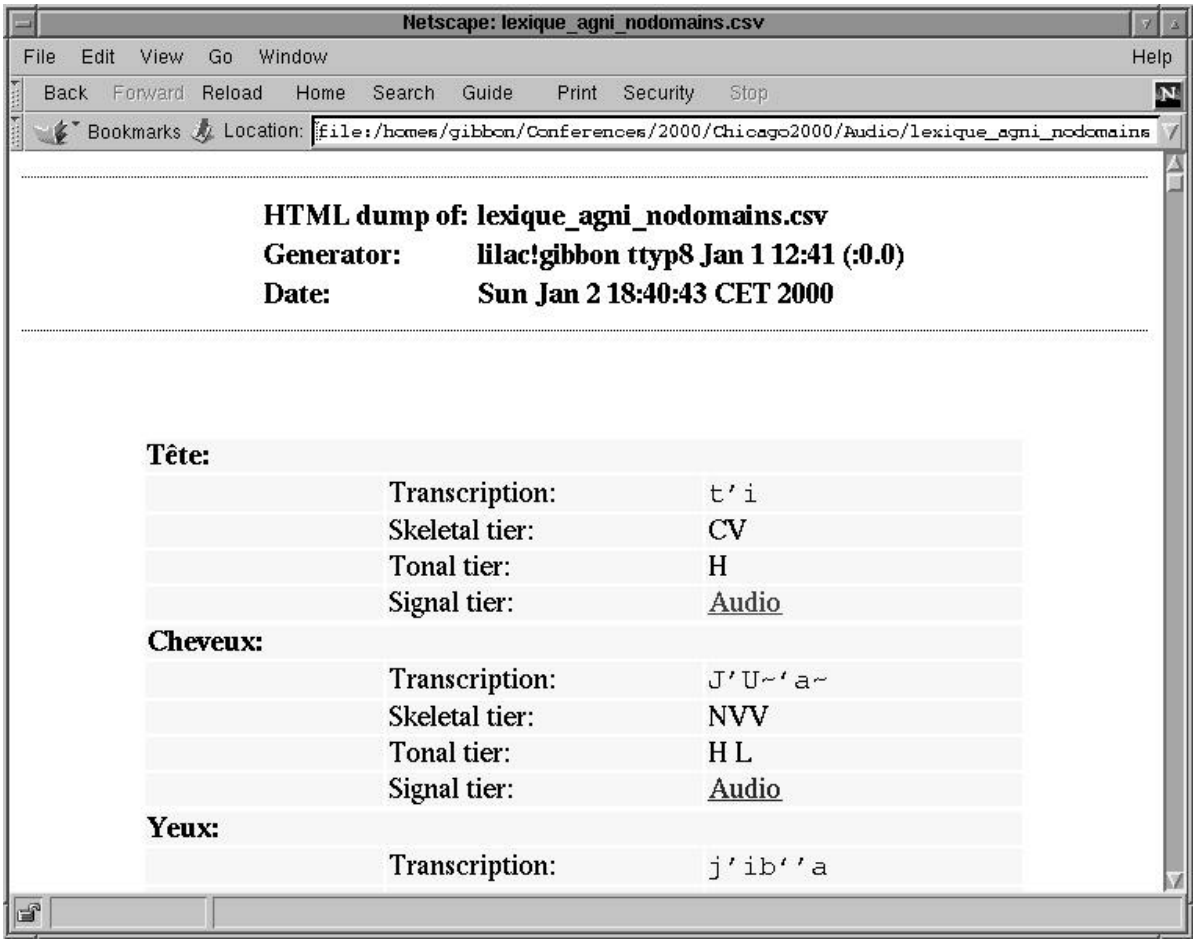
Implementation: Anyi database sample

Sample TAB-separated record:

‘an’U˜m‘’a˜ VNVNV L H LH Oiseau Agni_oiseau_1.wav

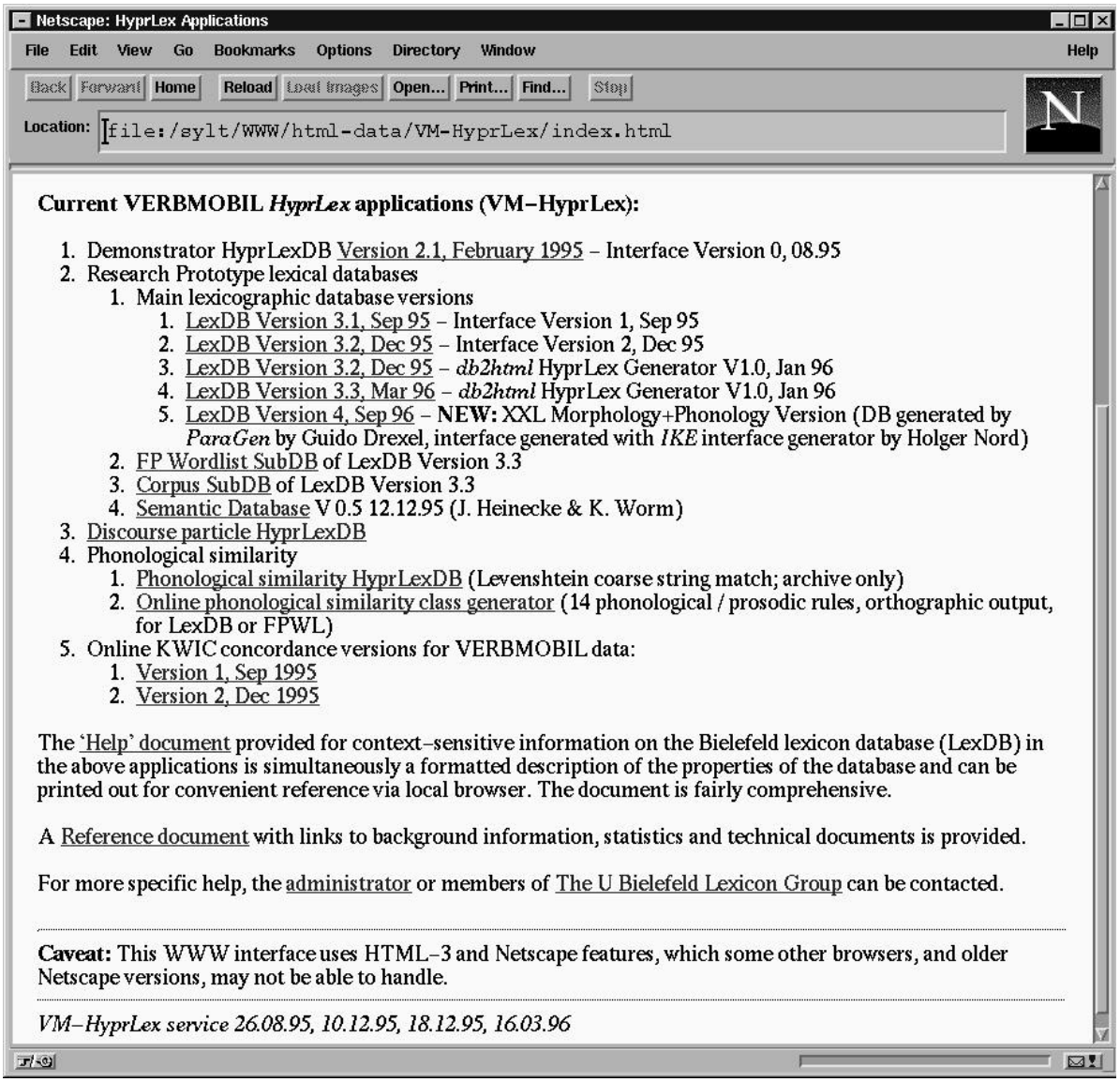
Transcription	Squelette	Ton	Glose	Fichier audio
t’i	CV	H	Tête	Agni_tete_1.wav
J’U˜‘a˜	NVV	H L	Cheveux	Agni_cheveux_1.wav
j’ib‘’a	CVCV	H LH	Yeux	Agni_yeux_1.wav
j’I˜s‘i	CVCV	H L	Sourcils	Agni_sourcils_1.wav
b’oJ	CVN	H	Nez	Agni_nez_1.wav
f’Ok’a	CVCV	H H	Joue	Agni_joue_1.wav
n’U˜‘a˜	NVV	H L	Bouche	Agni_bouche_1.wav
jé	CV	H	Dent	Agni_dent_1.wav
t‘ofl‘om’a˜	CVCLVNV	L L LH	Langue	Agni_langue_1.wav
w’U˜m’a˜	CVNV	H H	Front	Agni_front_1.wav
s’U	CV	H	Oreille	Agni_oreille_1.wav
k’Om’I˜	CVNV	H H	Cou	Agni_cou_1.wav
‘ah’ojè	VCVCV	L H L	Mâchoire	Agni_machoire_1.wav
s‘I	CV	L	Dos	Agni_dos_1.wav
k’U˜	CV	H	Ventre	Agni_ventre_1.wav
k‘Otw‘’a	CVCCV	L LH	Nombril	Agni_nombril_1.wav
b‘udr’E	CVCLV	L LH	Fesse	Agni_fesse_1.wav
k‘Ob‘’a	CVCV	L LH	Sexe féminin	Agni_sexefeminin_1.wav
tw‘’a	CCV	LH	Sexe masculin	Agni_sexemasculin_1.wav
b‘ut‘um’a˜	CVCVNV	L L LH	Anus	Agni_anus_1.wav
s’a	CV	H	Main	Agni_main_1.wav
s’ak’un-u	CVCVNV	H H M	Paume	Agni_paume_1.wav
s’ab‘’a	CVCV	H LH	Doigt	Agni_doigt_1.wav
s’ab‘u	CVCV	H L	Ongle	Agni_ongle_1.wav
jà	CV	L	Pied	Agni_pied_1.wav
jáb‘u	CVCV	H L	Orteil	Agni_orteil_1.wav
‘aw’Un’‘a˜J	VCVNVN	L H HL	Corps	Agni_corps_1.wav
pl’o	CLV	HL	Peau	Agni_peau_1.wav
b’ow’u‘e	CVCVV	H H L	Os	Agni_os_1.wav
m’ojà	NVCV	H L	Sang	Agni_sang_1.wav
‘em‘i’e	VNVV	L L H	Urine	Agni_urine_1.wav
n’U˜‘az˜’u‘e	NVVCVV	H L H L	Salive	Agni_salive_1.wav
j‘iz’u‘e	CVCVV	L H L	Larmes	Agni_larmes_1.wav
k’onv’i	CVNCV	H L	Gorge	Agni_gorge_1.wav
k’a˜z‘a	CVCV	LH	Menton	Agni_menton_1.wav
Jr’u	NLV	L H LH	Visage	Agni_visage_1.wav
‘ah’U˜b‘’a	VCVCV	H L	Coeur	Agni_coeur_1.wav
njófl‘a	NCVCLV	H L	Sein	Agni_sein_1.wav
k’a˜n’a˜	CVNV	H H	Plaie	Agni_plaie_1.wav
f‘a˜k’a	CVCV	L H	Muscle	Agni_muscle_1.wav

Implementation: LexDB HTML dump



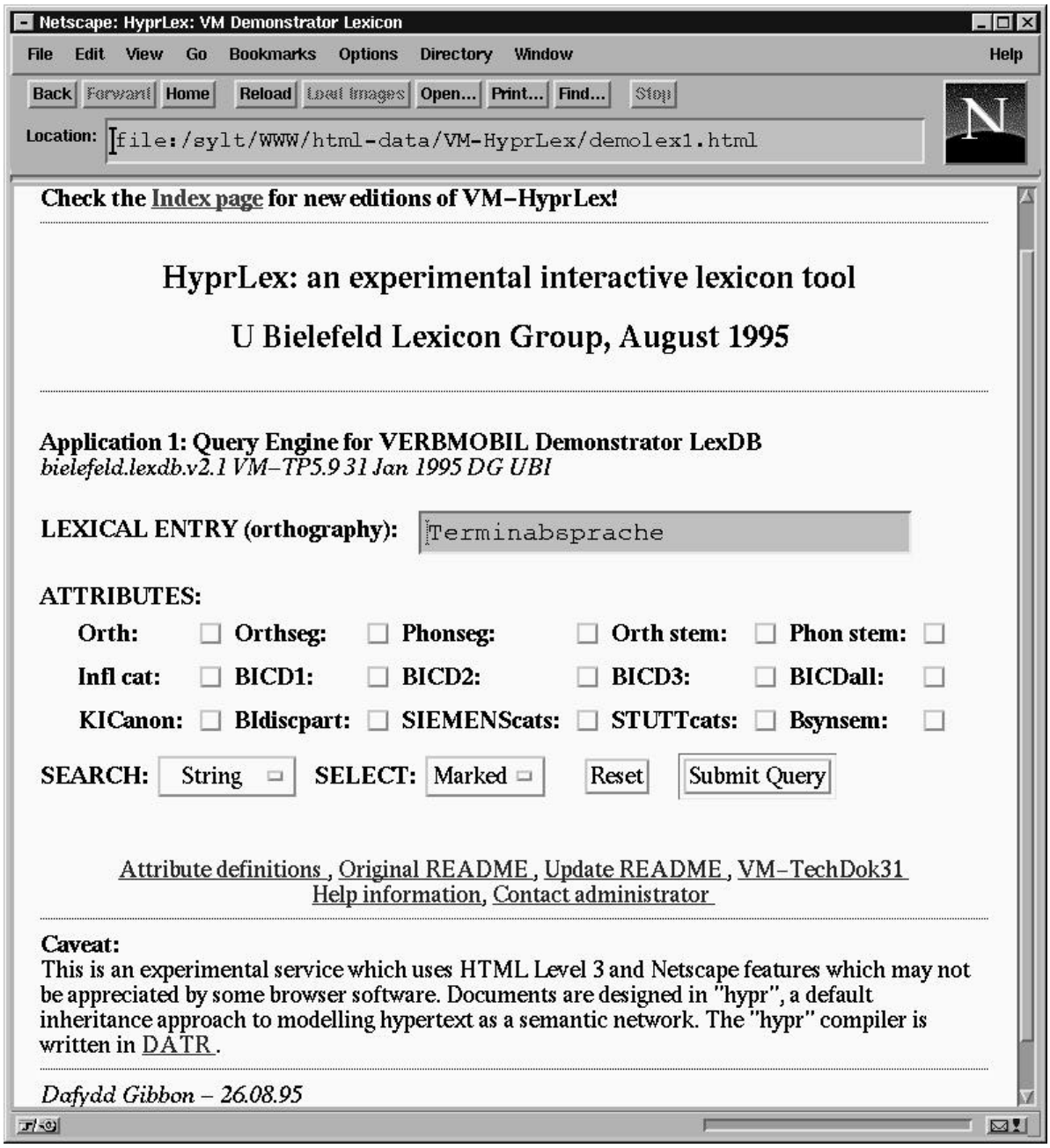
Implementation: the HyprLex LexDB family — 1

The VerbMobil HyprLex LexDB family - overview:



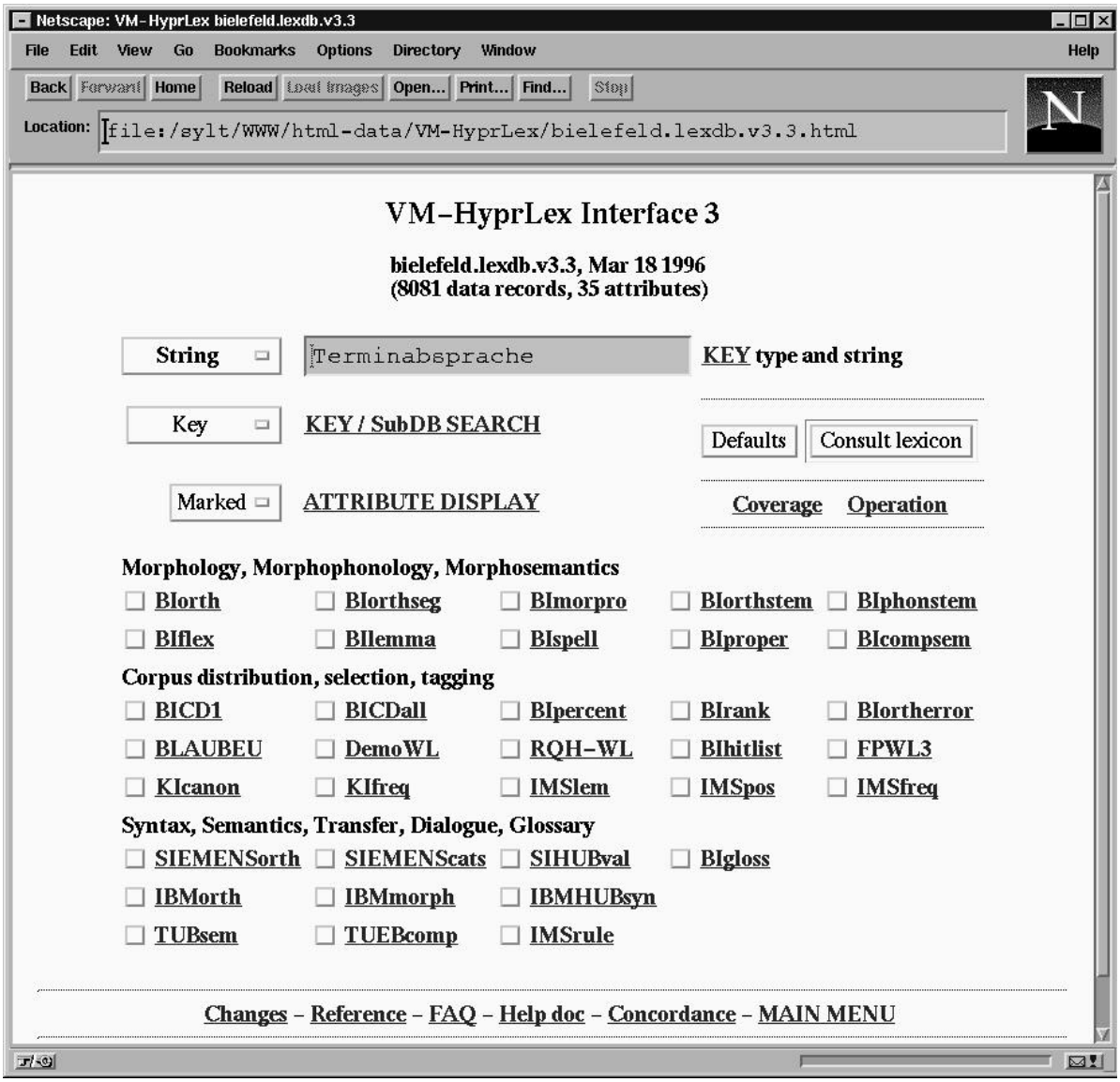
Implementation: the HyprLex LexDB family — 2

Beyond DTDs — 'Hypr': DATR as macro language for compiling families of HTML documents:



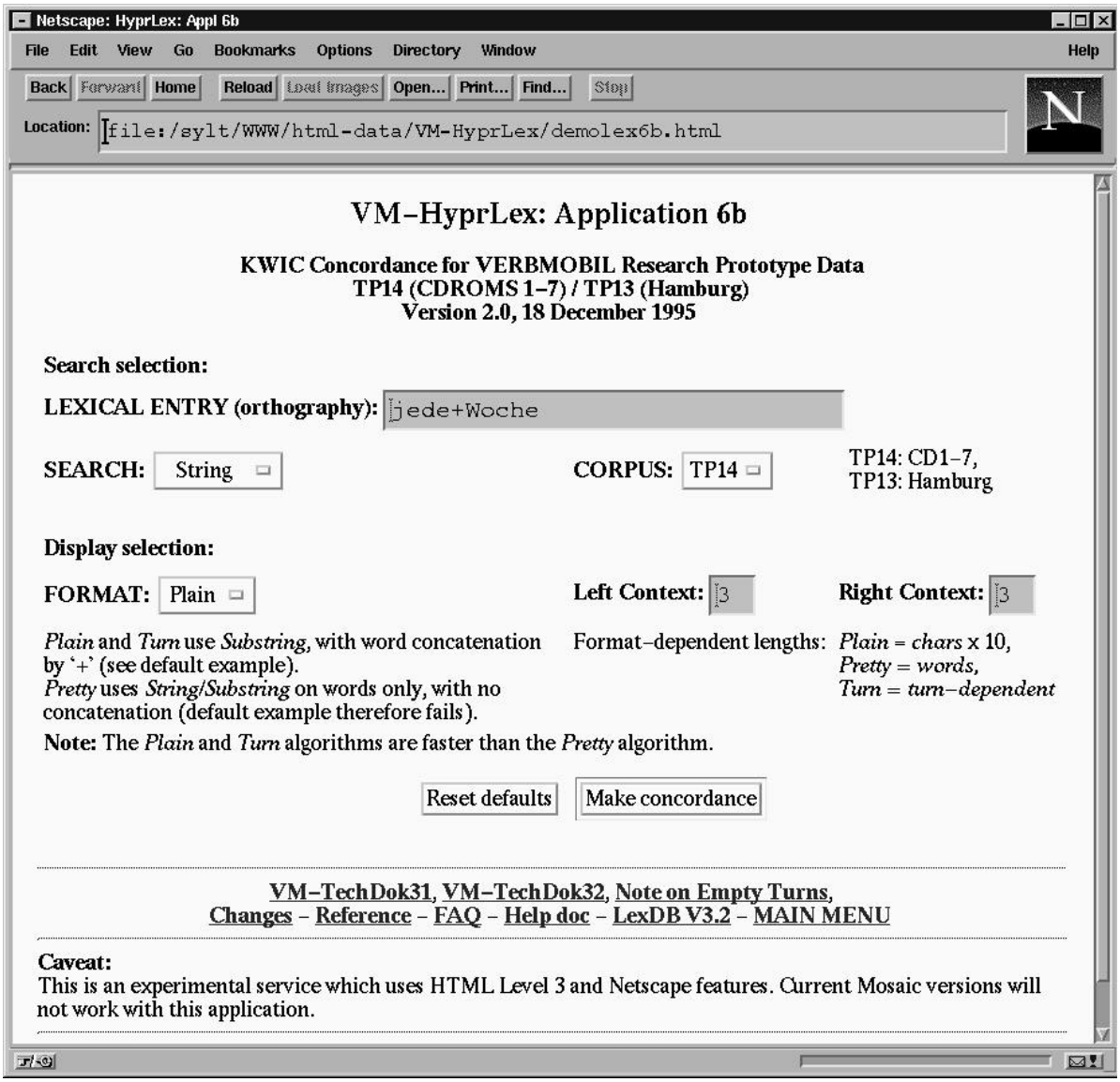
Implementation: the HyprLex LexDB family — 3

The main HyprLex LexDB:



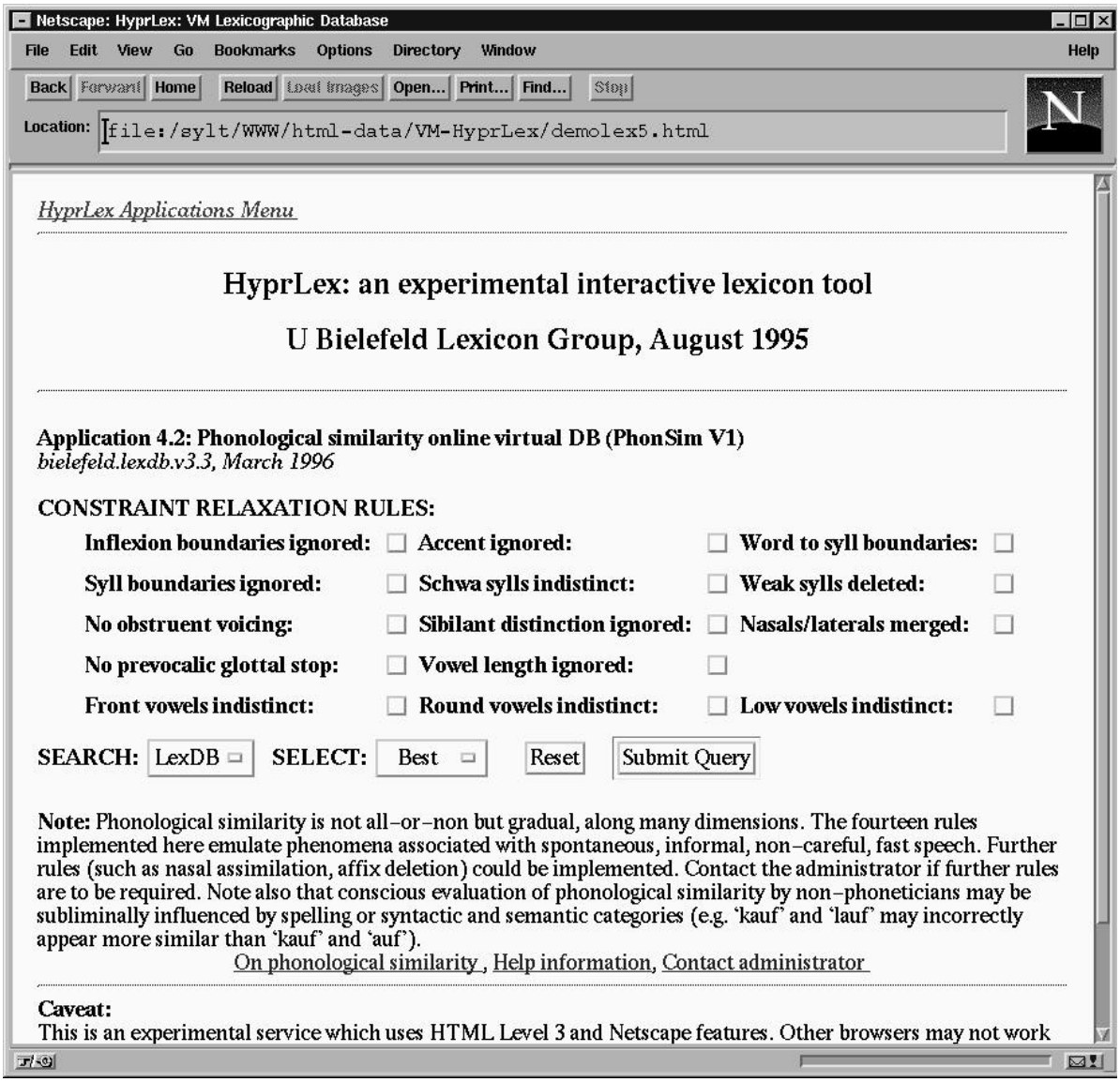
Implementation: the HyprLex LexDB — 4

The HyprLex KWIC concordance:

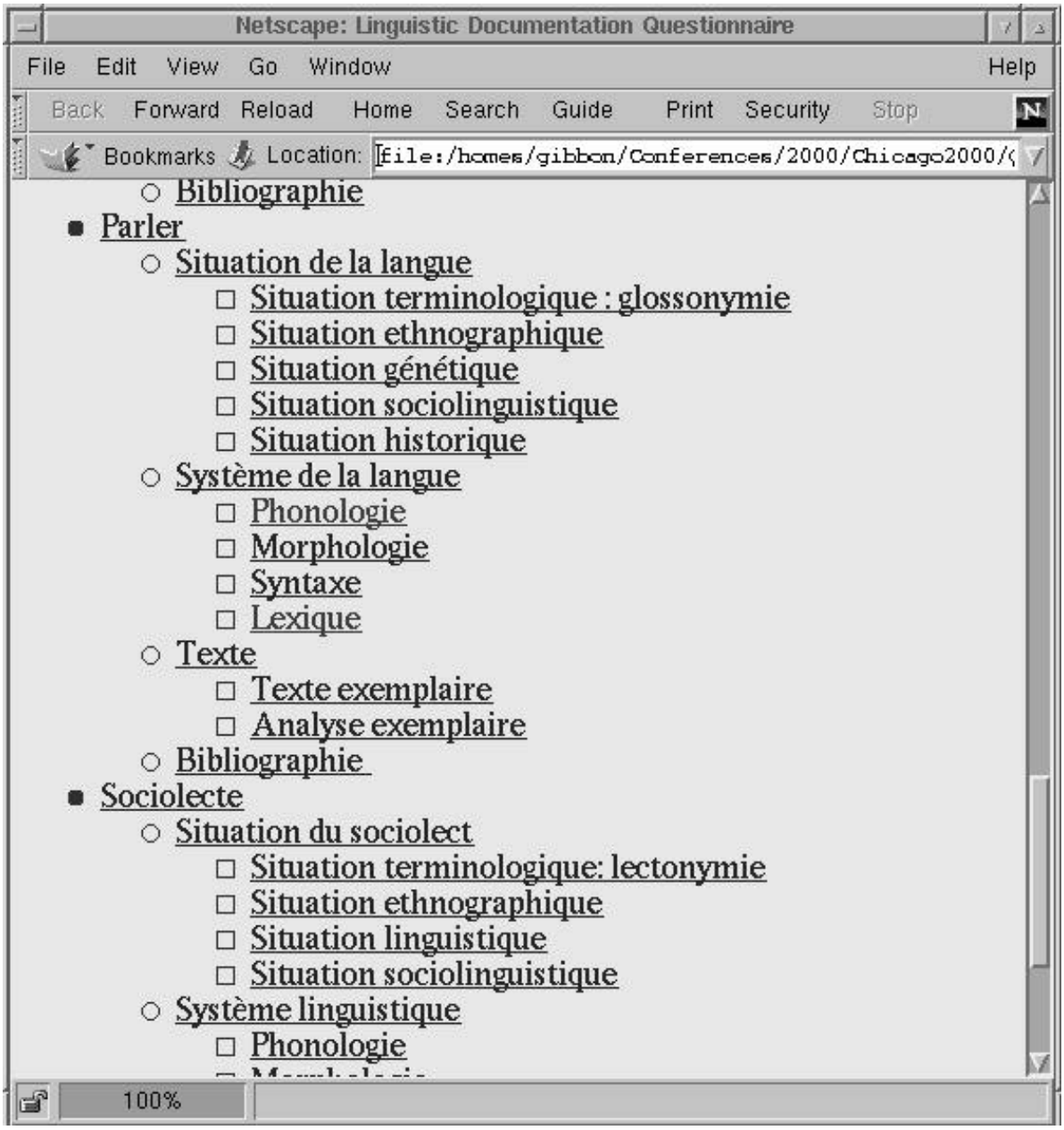


Implementation: the HyprLex LexDB — 5

The HyprLex phonological classifier:



Implementation: the hypertext questionnaire — 1



Implementation: the hypertext questionnaire — 2