ULex: new data models and a mobile environment for corpus enrichment

Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld

Postfach 1001312, 33739 Bielefeld, Germany

gibbon@uni-bielefeld.de

Abstract

The Ubiquitous Lexicon concept (ULex) has two sides. In the first kind of ubiquity, ULex combines prelexical corpus based lexicon extraction and formatting techniques from speech technology and corpus linguistics for both language documentation and basic speech technology (e.g. speech synthesis), and proposes new XML models for the basic datatypes concerned, in order to enable standardisastion and data interchange in these areas. The prelexical data types range from basic wordlists through diphone tables to concordance and interlinear glossing structures. While several proposals for standardising XML models of lexicon types are available, these more basic pre-lexical, data types, which are important in lexical acquisition, have received little attention. In the second area of ubiquity, ULex is implemented in a novel mobile environment to enable collaborative cross-platform use via a web application, either on the internet or, via a local hotspot, on an intranet, which runs not only on standard PC types but also on tablet computers and smartphones and is thereby also rendered truly ubiquitous in a geographical sense.

Keywords: computational lexicography, corpus enrichment, mobile computing

1. Objectives, motivation, requirements

1.1 Objectives and motivation

The ULex (Ubiquitous Lexicon) project has two main goals: first, the proposal of data models and standard XML format conventions for prelexical corpus enrichment in language documentation and speech and language technology; second, the provision of a generic platform-independent mobile online tool with a gentle learning curve, for extensive descriptive and technological language documentation. These data models and format conventions are used collaboratively via a web server on the internet or, via a local hotspot application, in local intranets.

Part of the motivation for ULex comes from extensive cooperation with field linguists who are used to linguistic and phonetic toolbox-type consumer software, and sometimes also to support from 'script hacking', and who do not initially realise the amount of practical support which systematic computational linguistic and natural language processing approaches can offer in terms of taking over mechanical distributional corpus processing tasks. The implementation is intended to demonstrate the potential of this kind of support as clearly as possible.

Another part of the motivation comes from the present lack of standard format specifications for prelexical data structures. One reason for this gap may be that these data structures are considered 'trivial', in some sense. Several such formats (e.g. concordances and interlinear glosses) are by no means trivial, however. In any case, simplicity is not a valid reason in itself for leaving such data structures in non-interoperable formats which require opportunistic local measures such as *ad hoc* scripting to be undertaken.

1.2 Prelexical data models

Some important corpus-based prelexical data types, i.e. data types involved in acquiring lexical data while creating a lexicon, are not covered in the ISO/TC-37/SC4 Lexicon Markup Framework (LMF) ISO 24613:2008 standard. Corpus enrichment with prelexical data structures includes not only commonly used formats such as grapheme-phoneme parallel text conversion, sorted word lists, frequency lists, pronunciation lexicons, KeyWord In Context (KWIC) concordances and InterLinear Gloss (ILG) annotation, but also, for example, diphone and other unit lists for speech synthesis. Interoperable and sustainable corpus enrichment with such prelexical data structures is essential both in documentary linguistics and in system development for human language technologies.

1.3 The ULex online tool

The platform independent Ubiquitous Lexicon (ULex) mobile online tool is an acquisition environment designed for creating enriched prelexical data in three main overlapping use cases:

(1) Proof-of-concept demonstration of markup techniques for prelexical data structures.

(2) Corpus checking and proof-reading in a 'corpus-cleaning' cycle, using sorted lists as an aid in removing user input errors.

(3) Minimising automatable mechanical tasks in corpus enrichment for lexicon acquisition.

(4) Provision of a tool for use in isolated intranets.

(5) Language documentation teaching.

The tool is intended for computationally relatively unversed users, and is not intended to provide NLTK-type comprehensive text mining faciliies (cf. Bird, Klein & Loper 2010).

The fourth case needs more clarification. The ULex tool should be ubiquitous, i.e. cross-disciplinary, platform independent (from workstations to smartphones) and location independent, with deployment potential for field linguistics and speech technology, especially in areas with little or no Internet connectivity.

Some components of ULex have simpler predecessors in previous projects (cf. Lüngen & Gibbon 2000; van Eynde & Gibbon 2000). The tool is currently being tested in field linguistics and speech technology for African and Central and South Asian languages.

The ubiquity promise in a similar sense has been around for some time (Gibbon 2002), but has never been really fulfilled. However, modern mobile devices make such specifications achievable (cf. Figure 5).

Additionally, the tool should have a gentle learning curve, with a simple 'one click' type interface. There are popular 'Swiss army knife' lexical tools with some prelexical functionalities, such as SIL's Toolbox and Fieldworks (SIL International, n.d.). However, they are platform specific, have notoriously steep learning curves, and do not yield some of the needed prelexical data structures. Other tools are dependent on Internet 'cloud services', and are consequently entirely non -interoperable in areas with poor digital infrastructure. Finally, the tool should be generic, language-independent (with language-specific code-tables), and wherever possible the tool should be standards conformant.

1.4 Overview

The following sections outline specific design decisions and implementation steps for attaining these goals, including design principles pertaining to functionality, constraints and user orientation, character tables and data models, and the ULex mobile client-server tool. The final section discusses standards conformity, evaluation, and prospects for further development. Owing to formatting constraints, the five Figures are listed at the end of the paper.

2. ULex design principles

2.1 Functionality

ULex is designed for both corpus-proofreading and corpus enrichment tasks. The specific design features are as follows:

1. Adherence to standards where possible (Unicode in UTF-8 encoding, XML, OLAC, ISO), but but definitions of new data models for pre-lexical data structures and of XML implementations for these data models are introduced where needed.
2. Definition of 13 prelexical data sstructures and 3 formats for enriched corpus output (i.e. a total of 39 options overall):
    1. grapheme-phoneme conversion (transliter-ation, phonetisation) of text;
    2. word lists (token list, type frequency list, pronunciation dictionary, KWIC concordance);
    3. interlinear gloss models (with 2 differently formalised gloss linking techniques);
    4. digraph and diphone lists (each with attested and potential combinatorics and frequency list).
3. Creation of an operational model: mobile client-server architecture, onboard HTTP server, wireless (less usefully: wired) clients.

For the user, the workflow of the tool is deliberately very simple: on the main corpus entry page a text is (currently) pasted into the entry area, parameter selections are made, and output is emitted to a web page with formatted or plain text output, and/or a download link. The three selectable interface parameters, for which drop-down select lists are prov ided, are: corpus (code-table selection for phonetisation mapping); data structure creation procedure (for the data structured listed above); output format (screen readable tables; XML; CSV).

2.2 Interoperability

The general design goal is to use standard data models for interoperability. In cases where no standards exist, as with prelexical data structures, this goal cannot be taken literally, but new data models need to be developed and proposed as potential standards.

For this reason, models for prelexical data structures are defined. As far as possible the definitions are constructed by analogy on the one hand with existing lexicon standards (ISO-24613:2008 LMF), and on the other hand with existing corpus resource model proposals (Hughes, Bird & Bow 2003).

2.3 Constraints and user orientation

The input corpus for the ULex tool should be orthographically as homogeneous and free from character and format ambiguities and inconsistencies as possible. No statistical generalisation or 'noise-checking' is provided at this stage of development. However, the ULex word list outputs are designed specifically to aid in 'corpus cleaning', i.e. in the necessary corpus checking and correction cycle, by consulting sorted lists of corpus units. Users are sometimes unaware of their own varied contributions to the inevitable need for such a cycle, for example in the form of inconsistent 'noisy' and ambiguous corpus codings, and experience shows that users need to be trained to be patient in this respect, steepening the learning curve somewhat.

The phonetisation process can be complex, a well-known fact both in speech technology and in linguistics. The basic constraint is that it is only possible to use simple character translation tables for orthographies with a biunique grapheme-phoneme relation. Newer orthographies often fulfil (or nearly fulfil) this requirement. However, well-established older orthographic systems (extreme cases are French and English) in general do not have this property, but have phonetisation variants which result from historical sound changes, which are dependent on morphological analysis, or which are loan words with their alien original spellings.

Fortunately, cases of linear dependencies and partial irregularities in grapheme-phoneme relations can in general be modelled as regular relations and implemented by means of Finite State Transducers, which can in turn be represented as ordered tables of character sequences where (for example) alphabetic ordering is used, and where grapheme sequences are ordered longest first, effectively yielding a logical default-override relation. This kind of table models a Deterministic Finite State Transducer (DFST; also 'Deterministic Finite Machine'), which is known to be an adequate representation for segmental (i.e. non-cyclical) phonological rules (Beesley & Karttunen 2003).

2.4 Corpus

The ULex proof-of-concept demonstrator contains a small corpus of Uyghur, a well documented but digitally under-resourced Turkic language of Western China (ISO 639-2: *uig*); 8 million speakers) with several scripts, including Roman, Perso-Arabic, Russian and Chinese; A Uyghur Latin Yëziqi (ULY) corpus is used here. Corpora and code tables for other African and Central and South Asian languages are undergoing testing. The demonstration corpus consists mainly of the Uyghur version of the UN Declaration of Human Rights, which is freely available on the Internet.

2.5 Character tables

The basic Latin UTF-8 encoded Unicode characters and common IPA phonetic characters are pre-defined, so that character translation tables can be sparsely populated: the user only specifies special characters from other code blocks, including less common IPA symbols, resulting in considerable labour-saving. Upper case Basic Latin characters, and some common characters from European alphabets, are handled automatically, but other upper case characters need explicit conversion. The sparse ULY character transformation table (Table 1) follows Saimaiti & Feng (2007). Engesæth & al. (2010) and Wikipedia (2011) entries on the Uyghur language were also consulted. The common substitutions "É, é" for "Ë, ë" are included. The Common Gateway Interface (CGI) server requires explicit UTF-8 encoding and decoding.

ULY	CGI UTF-8	IPA	U-codepoint
.		‖	
,		|	
-		(sp)	
'		ʔ	0294
a		ɑ	0251
e		æ	00E6
j		d͡ʒ	0361 0292
ch		t͡ʃ	02 0361 0283
zh		ʒ	0292
sh		ʃ	0283
gh		ʁ	0281
g		ŋ	014B
Ö	C3 96	ø	00F8
ö	C3 B6	ø	00F8
Ü	C3 9C	y	
ü	C2 BC	y	
w		v	
Ë	C3 8B	e	
ë	C3 AB	e	
É		e	
é		e	
i		ɪ	026A
y		j	

Table 1: Sparsely populated ULY (Uyghur) UTF-8 and Unicode codepoint character table.

2.6 XML models for ILG data structures

Hughes, Bird & Bow (2003) proposed an InterLinear Gloss (ILG) model for the E-MELD project which concentrates on class hierarchies of the units used, but it does not handle the co-seriality (co-linear or precedence) relations and the parallelism of sibling

classes (such as graphemes and phonemes) in an ILG hierarchy.

Figure 1 shows XML representations of two ULex ILG models handling seriality and sibling linking: ULex A, where sibling ILG gloss instances are serially co-indexed individually, and ULex B, where sibling ILG-Pair instances are serially co-indexed together. The two (interconvertible) models are shown in XML representation in Figure 1. The LMF classes *FormRepresentation*, *Representation* and *TextRepresentation* are abbreviated for present purposes to *Orth* and *IPA*. As with the KWIC indexing concept, cross-serial dependencies between different parallel interlinear gloss tiers are preserved.

2.7 XML model for KWIC data structures

The proposed LMF-like ULex Keyword In Context (KWIC) data model is shown in XML format in Figure 2. The LMF header and *LexicalResource* class, with subordinate classes *GlobalInformation*, *Lexicon*, *LexicalEntry* and *Form* are retained.

KWIC models are not specified in the Lexical Markup Framework (LMF), so new classes are introduced to handle syntagmatic relations: *Keyword*, *Context*, *Focus* and *Subcontext*. Co-indexing is introduced in order to represent positional serial (linear precedence) relations of the keyword instance and its contexts (XML sibling classes are intrinsically unordered), and the indexing is preserved in order to indicate graphemic-phonemic cross-serial dependencies.

3. ULex mobile client-server tool

The ULex tool was initially implemented for demonstration and training purposes, with a simple 'one-click' user interface and choices of data model and output format implementations.

The user interface at the client side of the implementation is shown in Figure 3, which shows the top part of the ULex input page, with the input field and the selector buttons. The 'Procedure' menu is shown dropped down with the 13 options which were listed previously among the design decisions. The client interface runs on any standard browser, and has been tested on Firefox, Opera, and various WebKit implementations.

In the medium term, client-side implementations in Java or JavaScript will be introduced, but currently the ULex mobile server is implemented by server-side techniques using Unix/Linux scripting. The reason for this 'breadboard' style of implementation is to ensure portability and interoperability of the server between the many kinds of desktop and mobile Unix and Linux based devices, and to avoid possible incompatibilities between client devices.

Many different ULex client-server configurations are possible, and several have been tested, ranging from the classic Unix workstation plus arbitrary desktop operating system to configurations with both clients and server on mobile devices, and, indeed, with server and client on the same mobile device in localhost mode. A number of servers have been used, including Apache and Lighty ('lighttpd').

One ULex server-client configuration with mobile devices, showing the text input page served by the smartphone, is illustrated in Figure 5. The smartphone at top right runs the 'Lighty' web server under the webOS operating system (cf. Whitby 2009), which provides a standalone localhost client service, and simultaneously, via a mobile wireless hotspot application, a wireless service to a tablet via a local internet or intranet router (bottom right) and a laptop via the Internet (bottom left). The current implementation uses Palm/HP webOS but the necessary software is available for other Unix/Linux derived smartphone and tablet operating systems (e.g. maemo/meego, Android, iOS, the last two possibly involving non-standard 'jailbreaking'). Interoperability is particularly important with mobile devices, which currently have a rather fast turnover of operating systems and operating system versions. In fact phones running the webOS operating system are currently no longer being manufactured, though the operating system itself is being made available by HP in an open source version. Mobile wireless hotspot applications are increasingly becoming available for other operating systems. The application currently in use is freeTether v1.2.0 (Gaudet & Hope 2011), which enables the establishment of a local intranet independently of connections to the internet, e.g. via 3G systems, a functionality not available in all hotspot applications.

4. Conclusion: standards conformity, evaluation, prospects

Character encoding standards (Unicode, UTF-8) and ISO-639.3 language codes are used. XML markup is analogous to ISO-24613:2008 (LMF), with the KWIC and ILG classes as innovations. A CSV output format provides interoperability with traditional databases, spreadsheets and word processor tables. OLAC standard metadata (Bird & Simons 2001) are being introduced. The standard web server with CGI provides portability.

Formal quantitative evaluation is hardly possible with a system of this kind, though mechanical prelexical tasks are subjectively performed

immeasurably more efficiently and consistently with a single 'one click' tool than in traditional manual linguistic fashion or with a collection of separate ad hoc tools.

The new prelexical data models enhance interoperability and are hereby proposed as potential XML standards. Further, deployment of the ULex server on mobile devices results in physical ubiquity. An interesting result, in passing, was that non-technical users tended to see their own user errors indiscriminately as software 'bugs': whether input errors (typos, glyph confusions etc.), data coding errors and omissions, or unspecified expectations (e.g. underestimation of the complexity of grapheme-phoneme relations). For these reasons, users need some explicit instruction. Despite intensive use, system malfunctions, i.e. bugs proper, have not yet been detected in the system.

The ULex concept has brought the goal of ubiquity in corpus-based lexical acquisition closer. The next stage is to extend the UTF-8 and Unicode codepoint tables for other popular Unicode blocks, especially for IPA code tables. Versions (currently 0.2.10) are publicly accessible (with frequent but brief offline periods for maintenance) and from version 1.0.0 will be open sourced at the same address:

http://wwwhomes.uni-bielefeld.de/gibbon/ULex/

5. References

Beesley, K. and L. Karttunen (2003). *Finite State Morphology*. Stanford: CSLI Publications.

Bird, S., E. Klein and E. Loper (2010). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. 2nd edn. O'Reilly Media.

Bird, S. and G. Simons (2001). The OLAC metadata set and controlled vocabularies In: *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources*, pp. 7-18.

Engesæth, T, M. Yakup and A. Dwyer (2009). *TeklimakandinSalam: hazirqi zaman Uyghur tili qollanmisi / Greetings from the Teklimakan: a handbook of Modern Uyghur.* Lawrence, Kansas: University of Kansas ScholarWorks. ISBN 978-1-936153-03-9. PDF with streaming audio.

van Eynde, F. and D. Gibbon, eds. (2000). *Lexicon Development for Speech and Language Processing*. Kluwer, Dordrecht.

Hughes, B., S. Bird and C. Bow (2003). Encoding and Presenting Interlinear Text Using XML Technologies. Australasian Language Technology Workshop, December 10 2003.

Gaudet, E. and Ryan M. Hope (2011). Application: FreeTether. Version 1.2.0. http://www.webos-internals.org/wiki/Application:FreeTether (verified 2012-03-12).

Gibbon, Dafydd (2002). Ubiquitous multilingual corpus management in computational fieldwork, *Proceedings of LREC 2002, Las Palmas, Gran Canaria*.

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

Saimaiti M. and Feng, Z. (2007). A syllabification algorithm and syllable statistics of written Uyghur. *Proceedings of the 4th Corpus Linguistics Conference, Birmingham, UK*.

SIL International (n.d.). Field Linguist's Toolbox. http://www.sil.org/computing/toolbox/ (verified 2010-03-13).

Whitby, R. ed. (2009). WebOS Internals. http://www.webos-internals.org/wiki/Main_Page (verified 2010-03-13).

6. Acknowledgments

The work reported here owes much to discussions within a considerable number of project frameworks in particular with (in approximate chronological order): Thorsten Trippel Doris Bleiching and many others in the VerbMobil team; Eno-Abasi Urua, Moses Ekpenyong, Firmin Ahoua and the ABUILD team; Laurent Romary; Nicoletta Calzolari; Steven Bird, Baden Hughes and Catherine Bow; Helen Aristar Dry and the E-MELD and LEGO teams; Jolanta Bachan; Arienne Dwyer.

7. Figures

```
<ILG-Entry id="4">
        <Orth>
                <ILG-Gloss index="1"> uyqung</ILG-Gloss>
                <ILG-Gloss index="2"> yéter</ILG-Gloss>
                <ILG-Gloss index="3"> ,</ILG-Gloss>
        </Orth>
        <IPA>
                <ILG-Gloss index="1"> ujquŋ</ILG-Gloss>
                <ILG-Gloss index="2"> jetær</ILG-Gloss>
                <ILG-Gloss index="3"> |</ILG-Gloss>
        </IPA>
</ILG-Entry>
```

```
<ILG-Entry id="4">
        <ILG-Pair index="1">
                <Orth>uyqung</Orth>
                <IPA>ujquŋ</IPA>
        </ILG-Pair>
        <ILG-Pair index="2">
                <Orth>yéter</Orth>
                <IPA>jetær</IPA>
        </ILG-Pair>
        <ILG-Pair index="3">
                <Orth>,</Orth>
                <IPA>|</IPA>
        </ILG-Pair>
</ILG-Entry>
```

Figure 1: XML representations of two ILG models for hierarchically equal items. Left, ULex A: gloss linking by index in a complete gloss). Right, ULex B: gloss linking as XML siblings in an indexed pair.

```
<?xml version="1.0" encoding="UTF-8"?>
<LexicalResource dtdVersion="15">
  <GlobalInformation>
    <feat att="languageCoding" val="ISO 639-3">
  </GlobalInformation>
  <Lexicon type="kwic" sortorder="alpha" corpus="uig-uly">
  ...
    <LexicalEntry>
      <Form>
        <Keyword rank="26">bir</keyword>
  ...
        <Context serialno="5">
          <Subcontext rank="1"> ular</Subcontext>
          <Focus rank="1">bir</Focus>
          <Subcontext rank="2">lishiр</Subcontext>
          <Focus rank="2">bir</Focus>
          <Subcontext rank="3">qowm bolup</Subcontext>
        </Context>
        <Context serialno="6">
          <Subcontext rank="1"></Subcontext>
          <Focus rank="1">bir</Focus>
          <Subcontext rank="2">tilda sözlishidiken</Subcontext>
        </Context>
        <Context serialno="7">
          <Subcontext rank="1"> ularni</Subcontext>
          <Focus rank="1">bir</Focus>
          <Subcontext rank="2"></Subcontext>
        </Context>
  ...
      </Form>
    </LexicalEntry>
  ...
  </Lexicon>
</LexicalResource>
```

Figure 2: Example of proposed KWIC concordance XML format.

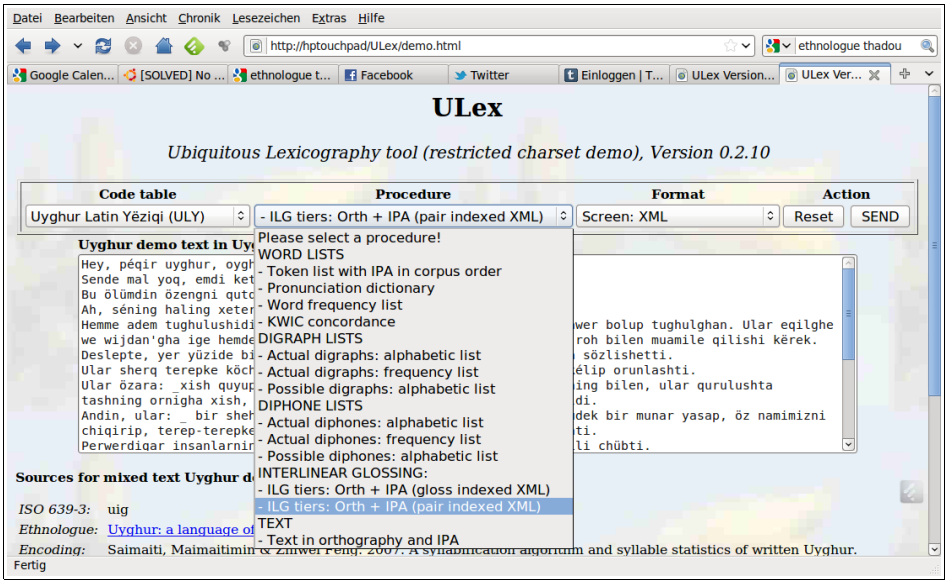

Figure 3: Top part of ULex input page showing data structure creation options.

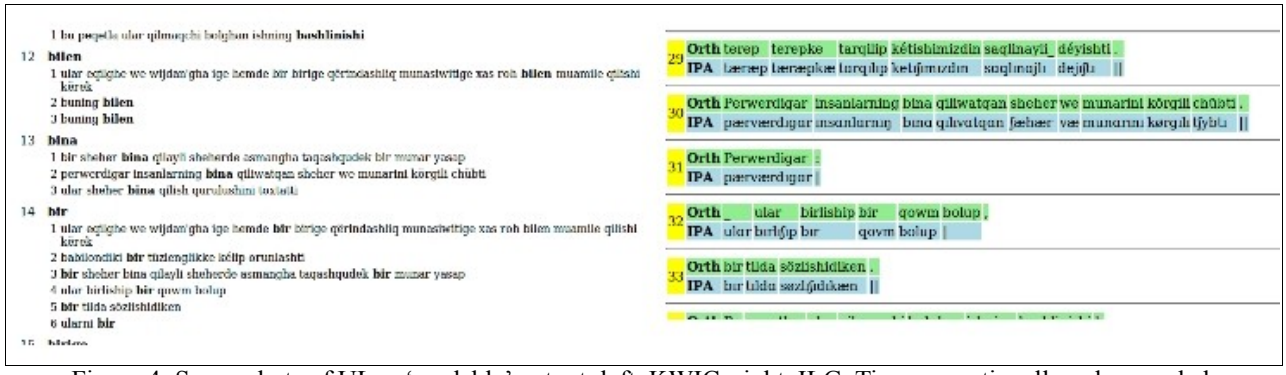

Figure 4: Screenshots of ULex 'readable' output: left, KWIC; right, ILG. Tiers are optionally colour-coded.

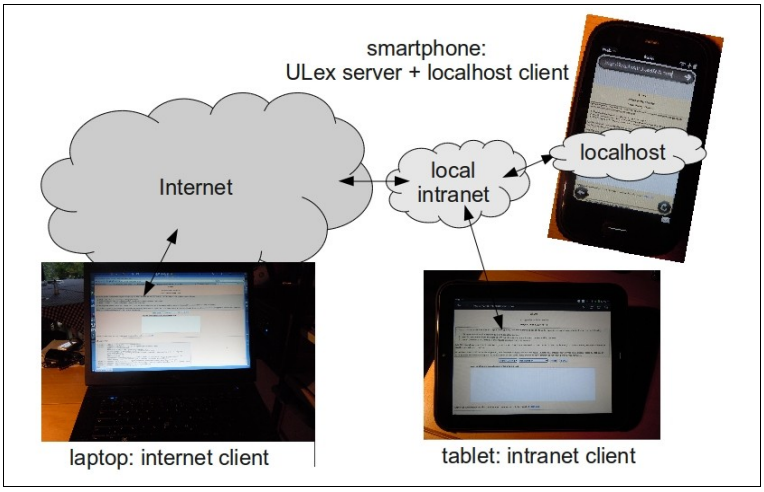

Figure 5: A client-server configuration with mobile devices.