Feature-based Encoding and Querying Language Resources with Character Semantics

Baden Hughes*, Dafydd Gibbon† and Thorsten Trippel†

*Department of Computer Science and Software Engineering
The University of Melbourne
Parkville VIC 3010, Australia
badenh@csse.unimelb.edu.au
†Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 100 131
D-33501 Bielefeld, Germany
{gibbon, thorsten.trippel}@uni-bielefeld.de

Abstract

In this paper we discuss the explicit representation of character features pertaining to written language resources, which we argue are critically necessary in the long term of archiving language data. Much focus on the creation of language resources and their associated preservation is at the level of the corpus itself; however it is generally accepted that long term interpretation of these language resources requires more than a best practice data format. In particular, where language resources are created in linguistic fieldwork, and especially for minority languages, the need for preservation not only of the resource itself, but of additional metadata which allows for the resource to be accurately interpreted in the future is becoming a topic of research in itself. In this paper we extend earlier work on semantically based character decomposition to include representation of character properties in a variety of models, and a mechanism for exploiting these properties through queries.

1. Introduction

In the context of archiving resources derived from linguistic fieldwork, particularly involving minority and endangered languages, it is insufficient to simply preserve the primary linguistic resources in isolation. In order to ensure long-term accessibility, a range of additional requirements must be met: not only must the data be stored in an application-independent format; information about the data itself must be preserved; the descriptive linguistic framework used to analyse and annotate the data must itself be documented.

The study of the relevant levels of descriptions for resources has become a research topic in iteself, strongly motivated by interoperability concerns in the context of long term archiving of language data. Perhaps most often expressed is the need for the descriptive properties implicit in a resource, by virtue of annotations of various types, to be grounded to higher level representations of linguistic meaning. While much recent work has focused on methods for expressing and integrating high level linguistic annotations, similar requirements exist at levels below the morposyntactic.

This paper describes the explicit representation of character features, which are necessary in the long term of archiving language data. While much focus on the creation of language resources (either large formal objects produced by agencies such as LDC or ELRA or smaller, dynamic resources produced by linguistic fieldwork) and their associated preservation is at the level of the corpus itself, it is well known that long term interpretation of these language resources requires more than a best practice format.

Specifically, in this paper we extend and exploit the semantic decomposition of characters into feature bundles in several different ways. First, we decompose a larger number of IPA characters into their feature vectors. Second, we consider the similarities between our decomposition and other standards for feature structures such as the ISO DIS 24610 Feature Structure Representation (International Standards Organization, 2006). Furthermore, we induce our feature vector model as XML, and thus enable common computational services to leverage the decomposition for higher level analyses. We show how various methods of enquiry are supported by the semantic character decomposition. Finally we consider a range of related work, draw a number of conclusions and identify directions for future work.

2. Background

In earlier work, (Gibbon et al., 2004) motivated consideration of the specific requirements for the archiving of language resources and their associated fonts, noting that for long term sustainability, simply archiving the font in which a minority language resource is represented is insufficient, and that the underlying semantics of the phonology/orthography relation are also required to secure interpretability.

At a high level, this paper draws motivation from (Bird and Simons, 2003), in particular the best practice recommendations for "*mapping the symbols used in transcription to phonological descriptiors which are mapped to a common ontology of linguistic terms*" and "*documenting any scheme used to transliterate characters*" (p. 575). It is in this light that we present our work on character transliteration schemes and associated analysis, which we label as the 'semantic decomposition of character encodings'.

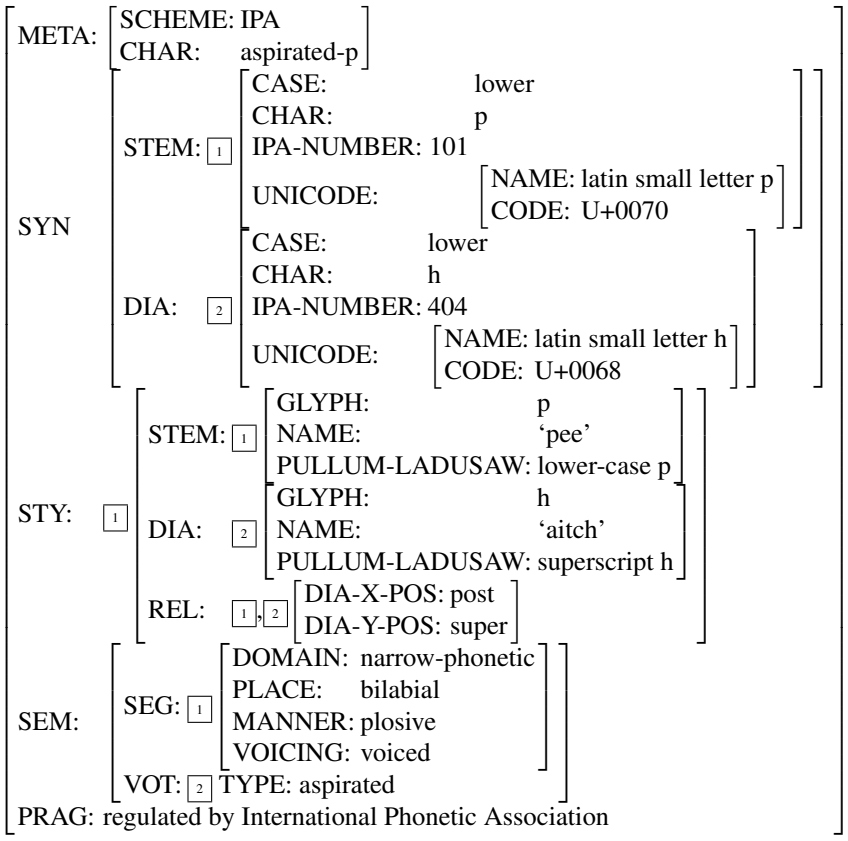

Figure 1: Structure of semiotic vector extract for [p^h] in IPA name-space.

Briefly revisiting the feature structure derived from our earlier work (Gibbon et al., 2006), we view a character as a composite of a number of different feature vectors for character syntax (**SYN**), character style semantics (**STY**) and character domain semantics (**SEM**); together with derivative feature sets for composition, glyph structure type, and sound type semantics. In doing so we integrate a number of sources which define feature hierarchies and oppositions for aspects of character encoding including the IPA (International Phonetic Association, 1999), Esling codes (Esling and Gaylord, 1993), Pullum and Ladusaw's phonetic symbol guide (Pullum and Ladusaw, 1986) and the Unicode standard (The Unicode Consortium, 2003). A sample decomposition along these lines is provided in Figure 1; we have since extended this decomposition across a broader section of the IPA, covering all vowels, a selection of fricatives, plosives and nasals[1]. Similarly, we have 'reverse engineered' the character semantics for resources in the Ega language, which can be found at an alternative location[2].

3. A Feature Structure Representation

Having reviewed the decomposition formalism the astute reader may notice similarities between our graphical feature structure representation and the 'matrix' model proposed by ISO-DIS-24610-1. As far as we are aware, this is one of the first non-morphosyntactic uses of the ISO-DIS-24610-1 Feature Structure Standard. This is fortuitous for numerous reasons; including providing a core framework for data integration with other linguistic data types. It is in fact feasible to construe our semiotic vector matrix as any of the models provided within the ISO-DIS-24610-1 Feature Structure Representation. The advantage of the underlying semantic decomposition is that each feature structure is both data-typed and constrained. From these closed value classes, we can induce a Document Grammar eg. a DTD or XML Schema, allowing a syntactic parser to validate a character description (or set of such descriptions). By the extensive use of closed classes and data types in newer schema languages, the overall quality of the resource description can also be increased significantly.

4. An XML Expression

Typed feature structures such as our semantic decomposition have a well-recognised natural representation in XML; which we can exploit independently of a formal feature structure representation such as ISO-DIS-24610-1. An XML expression corresponding to the feature structure in Figure 1 can be seen in Figure 2.

The main issues associated with rendering our feature structure in XML are in the area of the type of expression which is most suitable for the properties expressed. Fortunately, our feature structure is a 3 level hierarchy, which naturally lends itself to the `element:attribute:value` triple which is the default in an XML representation.

Naturally a corresponding Document Type Definition and an XML Schema (World Wide Web Consortium, 2001)

[1] See `http://www.spectrum.uni-bielefeld.de/LangDoc/charsem`

[2] See `http://www.emeld.org/school/case/ega`

```
<?xml version="1.0" encoding="utf-8"?>
<character>
 <meta scheme="ipa" char="aspirated-p"/>
 <syn>
  <stem object="1" case="lower" char="p" ipa_number="101"
   unicode_name="latin small letter p" unicode_code="U+0070"/>
  <dia object="2" case="lower" char="h" ipa_number="404"
   unicode_name="latin small letter h" unicode_code="U+0068"/>
 </syn>
 <sty>
  <stem object="1" glyph="p" name="pee" pullum-ladusaw="lower-case p"/>
  <dia object="2" glyph="h" name="aitch" pullum-ladusaw="superscript h"/>
  <rel object="1,2" dia-x-pos="post" dia-y-pos="super"/>
 </sty>
 <sem>
  <seg object="1" domain="narrow-phonetic" place="bilabial" manner="plosive"
   voicing="voiced"/>
  <vot object="2" type="aspirated"/>
 </sem>
 <prag comment="regulated by the International Phonetic Association"/>
</character>
</xml>
```

Figure 2: XML encoding [p^h] in IPA name-space.

or RELAX-NG schema (International Standards Organization, 2002) is required to facilitate computational processing of such an XML based expression. We have developed the DTD and XML Schema (see the URL reported earlier in the paper), and are working on a RELAX-NG schema.

5. Transformations

We can trivially translate our XML-encoded feature structure to the forms prescribed in ISO-DIS-24610-1 (including the graph, matrix or XML variants). Furthermore, we can support directly the ISO-DIS-24610-1 concept of collections for complex feature bundles, demonstrating the use of such mechanisms beyond morpho-syntax.

In addition, our representation is trivially decomposed into a relational database schema for implementation in SQL-based infrastructure; or into a state based automaton such as the finite state models proposed by (Garrett, 2005). For query purposes, a rendition of this data into one of these forms is likely to increase computational tractability at an application level.

Furthermore, XSL Transformations which are inherently display oriented can render such a data structure for human consumption (eg through a web browser). Rendering the character set in the model form of the IPA is conceivable as one such demonstration. Conversely XSLT can be used to transform the XML representation to a different structure (eg for input to a different system such as a reasoner in Prolog or a lexical inference engine in DATR).

By extension, any language with support for XML parsing can also transform the underlying XML representation into a different form. Our early work has been with Python and Perl, although increasingly Java is the language of choice for such manipulations where XSLT and XQuery are insufficient.

6. Querying Encoded Resources

Semantically grounded phonetic and phonological enquiry can leverage this encoding to query written language resources. In particular we can execute queries on a number of different levels of the character semantics including at structural (through unit types such as symbols, diacritics, sequences); semantic (phonetic feature structures based on the semantics of the IPA as represented in the standard IPA chart) and representational (visual through fonts, codes, glyphs; and acoustic through mappings to signals).

In earlier work (Gibbon et al., 2006) we posited a number of preliminary query cases for exploring characters semantics; exploiting the representational mechanism's encoding of various properties which remain hidden at a text level. In particular we motivated exploration across multiple dimensions including including their proximity in the semiotic vector space, in linguistic meaning, structure and context-sensitive rendering, provenance throughout a family of related fonts etc. Here we continue this theme by proposing prototypical queries, their expression in an XML query language, and their results. In particular we focus on mining similarities between characters determined by generalisations (attribute–value structure intersections) over feature structures at various hierarchical levels:

SYN: UNICODE values (or other font or encoding values such as ASCII, SILDOULOS);
CASE, CHAR, CODE values (by further decomposition on Unicode principles);
STEM, DIACRITIC values;

STY: GLYPH, HOR–POS, Y–POS values;
GLYPH STATUS, DIACRITIC values;

SEM: DOMAIN, PLACE, MANNER, VOICING values;

– SEGMENT, VOICEONSET values;

META: CHAR; SCHEME;

PRAG: regulatory criteria and versioning; definitions of orthographic and phonemic coverage of a given language.

The classification task in this context is relatively straightforward, since for most cases the questions will be related to the similarity or differences of a given character or font. In our more formal context, we can not only identify the differences, but quantitify them and ground them in a domain of interpretation. This represents a significant advancement over the ad hoc, manual inspection methods which currently characterise the field of comparative linguistic encoding analysis.

As far as a query model is concerned, there is a high degree of flexibility, although in keeping with traditional practice, query constraints generally are expressed at the most coarse grained level and then refined to finer grained distinctions.

Example queries based on the semantic decomposition formalism could reasonably include the following:

- finding the set of Unicode code points used in the digital encoding of a given language resource,
- locating one or more fonts with the best coverage of the characters needed for a given language based on properties such as the presence of diacritics,
- identifying phonetic or phonological features of interest across various matrices derived from written rather than spoken language data
- exploring the nexus between orthographic inventory and phonetic inventory for a given language
- finding the IPA attribute-value matrix of phonetic features of a given language

Relational query languages such as SQL, semistructured data query languages such as XQuery (Boag et al., 2005), and linguistic query languages such as LPath (Bird et al., 2005) are natural partners for the exploitation of the XML representation of the semantic properties of characters.

7. Related Work

It should be noted at the outset that our approach is indicative of a theme of research which has been strongly motivated by the requirements of field linguists for data mining operations over primary language data. We can establish the procedural similarities in feature based exploration between our work and others such as HyperLex (Bird, 1999a) and HyprLex (Gibbon and Trippel, 2000); related activity in the areas other phonological decompositions motivated by query requirements (Garrett, 2005); and to similar feature based decompositions for exploring language change (Kamholz, 2005).

In HyperLex (Bird, 1999a), a similar decomposition and enquiry approach was demonstrated with regard to complex tone properties in Bamileke Dschang (a West African language). In HyperLex, tone is explored through the paradigm of feature bundles which are both lexically and phonologically based. Feature bundles are construed as n-dimensional data cubes, which can be sliced according to various features of interest.

More evidence for the viability of this approach is provided by (Gibbon and Trippel, 2000), through their HyprLex system. Although focused on the lexicon, as opposed to the characters, they demonstrate a similarly-principled semiotic vector decomposition of both implicit and explicit properties into feature bundles which form the locus of later enquiry. A highly tractable computational expression was derived using DATR.

In other related work, (Garrett, 2005) describes the interpretation of phonetic feature values as a finite state representation. He proposes a finite state transducer that converts text represented in the International Phonetic Alphabet into phonetic feature sets, or vice versa; accounting for complex decomposition issues induced by Unicode. A system has already been inplemented which supports this work[3]. Our work here has much in common with his approach; although a careful review of this work shows there are some advantages to our approach (namely, not being bound to normalized and decomposed Unicode representations).

Motivated by the need to track phonetic changes in the broader context of language change, (Kamholz, 2005) has also provided a semantically motivated decomposition of IPA characters. This work has its genesis in the idea of evolutionary phonology (Blevins, 2004), and the need to regularise the properties of sounds in order to project and track diacrhonic change. Our work has similarities in the sense that the IPA provides the basis for some of the character properties, but differs by virtue of the additional properties we encode based on ontologies such as the Unicode character model, Pullum and Ladusaw etc. Notably, Kamholz has implemented a web based system for feature derivation from an IPA character based on a relational model of properties[4].

8. Future Work

A number of directions for future work can be identified.

Although we have made considerable progress on reducing the IPA attribute-value matrix into our formalism, this remains to be completed. In addition to IPA there are numerous other quasi-standards which are used to document the phonetic and phonological properties of a given language; most often expressed by the orthogragphic and font choice. We would like to decompose a range of other common fonts such as SIL's Doulos in the same manner.

Our original motivation for this work was that of securing the interpretability of minority language data which is endangered owing particularly to format, encoding or font obsolescence. In earlier work, we demonstrated the capacity of the semantically-based character decomposition

[3] http://altiplano.emich.edu/IPA4Unicode/

[4] http://brugmann.eva.mpg.de/~kamholz/phon.pl

in ensuring that legacy language resources are interpretable outside their original technological context, and we would like to perform similar analysis on other language resources of known levels of endangerment. Unfortunately, without considerable resourcing, this is likely to be an area of slow progress, but enormous gain.

In order to address the efficiency issue, there is a need for tools which both enable semantically-based character decomposition to be performed, and for application instances to leverage the resulting descriptions. We acknowldge that while suggesting to end users that this documentation form is valuable, there is unlikely to be large amounts of such documentation produced unless either primary documentation tools such as Shoebox/Toolbox or archive level services support this type of analysis as an integrated module within an overall workflow.

The potential for automated analysis of existing language data encoding schemes is significantly increased by the availability of formal descriptions of orthographic character properties. We can envisage the availability of tools which can consume a semantically based character decomposition set, and using this information to make assertions (either formal or informal) about the distribution of phonetic and phonological phenomena from written, rather than spoken language resources; and thus be used to correlate between written and spoken forms. This is particularly useful in context where there has been discontinuous interaction with a language by documentary linguists, or where archival data exists but is not well integrated into a descriptive framework.

On a related note, this mechanism provides us with the ability to consider issues at the boundary of the phonetic, phonological and orthographic boundaries. The intersection of these domains has been shown to have a range of interesting implications for the acquisition of literacy in a given language eg (Bird, 1999b), yet there is a lack of broad coverage data for exploring these impacts.

The availability of this data type also allows for a new type of language resource to be constructed: an integrated phonetic inventory and orthographic projection for a given language. It would be almost trivial to extend the XML formalism for to support a multimedia linkage (file, offset, duration)) for an exemplar of a given character, allowing for a richer description of a language.

Finally, there is need for language archives to support (and encourage) resources of this type. Since the base format is likely to be XML, no specific technical requirements are forthcoming, however metadata for these types of objects must be created, encessitating a consideration of how to decribe this resource type. In metdata frameworks such as the Open Language Archives Community, this type of resource would likely require a new descriptive type.

9. Conclusion

In this paper we have shown how the semantically-based decomposition of character properties allows for a new type of exploitation of written language resources. The procedure for deriving these properties from orthographic representations effectively reverse engineers the knowledge applied by linguists in reducing a language to a written form. In particular contexts, such as endangered language documentation, such round-trip approaches are invaluable in allowing the inherent linguistic analysis of a data set to be exposed for machine processing. By providing a decomposition model, a representation in a standard form, and example queries, we hope to provoke further interest in this type of language resource and linguistic enquiry.

10. References

Bird, S., 1999a. Multidimensional exploration of online linguistic field data. In *Proceedings of the 29th Meeting of the North-East Linguistic Society*.

Bird, S., Y. Chen, S. Davidson, H. Lee, and Y. Zheng, 2005. Extending XPath to Support Linguistic Queries. In *Proceedings of Programming Language Technologies for XML (PLANX)*.

Bird, S. and G. Simons, 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language*, 79(3):557–582.

Bird, Steven, 1999b. When marking tone reduces fluency: an orthography experiment in cameroon. *Language and Speech*, 42:83–115.

Blevins, Juliette, 2004. *Evolutionary Phonology: The emergence of sound patterns*. Cambridge University Press: Cambridge.

Boag, S., D. Chamberlin, M.F. Fernández, D. Florescu, J. Robie, and Jérôme Siméon, 2005. XML Query 1.0: An XML Query Language. http://www.w3.org/TR/xquery/.

Esling, J. H. and H. Gaylord, 1993. Computer codes for phonetic symbols. *Journal of the International Phonetic Association*, 23(2):83–97.

Garrett, E. J., 2005. A Finite State Network for Phonetic Text Processing. In *Computational Linguistics and Intelligent Text Processing, 6th International Conference, (CICLing 2005)*. Springer-Verlag: Berlin.

Gibbon, D., C. Bow, S. Bird, and B. Hughes, 2004. Securing Interpretability: The Case of Ega Language Documentation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. European Language Resources Association: Paris.

Gibbon, D., B. Hughes, and T. Trippel, 2006. Decomposition of Character Encodings for Linguistic Knowledge Discovery. In *Studies in Data Classification*, volume 30. Springer Verlag: Berlin.

Gibbon, D. and T. Trippel, 2000. A multi-view hyperlexicon resource for spoken language system development. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. European Language Resources Association: Paris.

International Phonetic Association, 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press: Cambridge.

International Standards Organization, 2002. *Document Schema Declaration Languages (DSDL) - Part 2: Regular-grammar-based validation - RELAX-NG (ISO/IEC FDIS 19757-2:2002(E))*. International Standards Organization: Geneva.

International Standards Organization, 2006. *Language resources management – Feature structures – Part 1: Feature structure representation ISO DIS 24610-1*. International Standards Organisation: Geneva.

Kamholz, D., 2005. An Ontology for Sounds and Sound Patterns. In *Proceedings of the EMELD Workshop on Digital Language Documentation: Linguistic Ontologies and Data Categories for Language Resources*. http://emeld.org/workshop/2005/papers/kamholz-paper.doc.

Pullum, G. K. and W. A. Ladusaw, 1986. *Phonetic Symbol Guide*. University of Chicago Press: Chicago.

The Unicode Consortium, 2003. *The Unicode Standard, Version 4.0*. Addison-Wesley.

World Wide Web Consortium, W3C, 2001. XML Schema. http://www.w3.org/XML/Schema.