Originally published in:Spiliopoulou, Myra/Kruse, Rudolf/Borgelt, Christian/Nürnberger, Andreas/ Gaul, Wolfgang (Eds.): From Data and Information Analysis to Knowledge Engineering. Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Magdeburg, March 9-11, 2005. – Berlin: Springer, 2006. Pp. 366-373.

POSTPRINT

Semantic Decomposition of Character Encodings for Linguistic Knowledge Discovery

Dafydd Gibbon[1], Baden Hughes[2], and Thorsten Trippel[1]

[1] Fakultät für Linguistik und Literaturwissenschaft
Universit¨at Bielefeld, Postfach 100 131, D–33501 Bielefeld, Germany

[2]Department of Computer Science and Software Engineering
University of Melbourne, Parkville 3010, Australia

Abstract
Analysis and knowledge representation of linguistic objects tends to focus on larger units (e.g. words) than print medium characters. We analyse characters as linguistic objects in their own right, with meaning, structure and form. Characters have meaning (the symbols of the International Phonetic Alphabet denote phonetic categories, the character represented by the glyph '∪' denotes set union), structure (they are composed of stems and parts such as descenders or diacritics or are ligatures), and form (they have a mapping to visual glyphs). Character encoding initatives such as Unicode tend to concentrate on the structure and form of characters and ignore their meaning in the sense discussed here. We suggest that our approach of including semantic decomposition and defining font–based namespaces for semantic character domains provides a long–term perspective of interoperability and tractability with regard to data–mining over characters by integrating information about characters into a coherent semiotically–based ontology. We demonstrate these principles in a case study of the International Phonetic Alphabet.

1 Introduction

High quality language documentation according to agreed professional standards is becoming an essential part of the empirical resources available for linguistic analysis, and a new subdiscipline, documentary linguistics, has emerged in this area [Himmelmann, 1998]. The main emphasis of the language documentation enterprise lies in three areas: the provision of extensive and consistently annotated development data for the human language technologies, the sustainable and interpretable preservation of endangered languages data [Gibbon et al, 2004] and the professional archiving of documents of any kind by the methods of text technology.

In contrast, little attention has been paid from a linguistic point of view to the incorporation of the smallest structural units of written texts, characters, into this enterprise. On closer inspection, characters, character sets and encodings which are used to represent textual data turn out to be a linguistic domain in their own right, but one which has hardly been explored.

Our contribution is to introduce a new approach to character decomposition and classification, and an outline formalisation of this approach. First we discuss encoding strategies, from legacy practice through current Unicode

practice to the need for a more generic approach. We provide a case study around the International Phonetic Alphabet, defining characters as linguistic signs, and examining their properties according to a linguistic model which relates meaning, structure and form, with properties represented as feature vectors or attribute–value matrices (AVMs) according to current notational conventions in general and computational linguistics.

The generic character descriptions are used to explicate conventional Unicode and non–Unicode character encodings. We then show how semantic character decomposition brings advantages for the representation of user–oriented properties of characters, such as their linguistic meanings, their structures, or their context–sensitive rendering. In order to show how to overcome problems of missing characters in typical uses we discuss an ontological approach to character mapping, based on the idea of fonts as namespaces with mappings to a variety of encodings.

The domain of character encoding has a number of importantly differentiated terms and concepts which are often employed loosely in everyday use, e.g. *character*, *letter*, *text element*, and *glyph*. These terms need to be clearly defined in order to appreciate the context of the remainder of this work. We proceed after the model of [Dürst et al, 2004] and [Unicode Consortium, 2003], in defining a character, its various renderings, and the text processes, input methods, collation approaches and storage requirements; these sources should be consulted for further detail on encoding.

In addition to the character, its rendering and its role in text processes we are also interested in the semantics and pragmatics of characters, i.e. the meaning and role of characters in the usage contexts of language communities, and in the development of a generic classification of characters from this point of view in a coherent and comprehensive character ontology.

We avoid both glyph–based 'lookalike' and code–based criteria, and take a linguistic approach to solving the problem of unifying the linguistic properties of both best–practice and legacy character encodings. We have developed an analytical, classificatory and representational approach independent of specific fonts or character encodings, and at a higher generic level than Unicode, in that provision is made for including coherent user–oriented semantics and pragmatics of characters. The representational meta–syntax we use is attribute–value based; for applications in interchange and archiving there is a straightforward mapping into the more verbose XML notational conventions.

2 Characters as Signs: A Case Study of IPA Characters

The body of this work is a short case study of an application area for semantic character decomposition in which feature–based character descriptions are developed as the basic units of a character ontology for character–based data–mining tasks in the context of the semantic web.

We define a character as a linguistic sign and decompose its semantics into linguistic feature vectors representing semantic interpretation (in phonetic, phonemic and orthographic worlds), structure, and glyph rendering interpretation. The decomposition inherits a range of properties from Unicode concepts such as inherent directionality and combining behaviour, and the result is applicable both to Unicode and non–Unicode character encodings.

A commonly used standard character set is the International Phonetic Alphabet (IPA). The standardizing body is the International Phonetic Association, which periodically considers revisions to the character set. The organisation of the properties of characters in this set may be expressed as a vector $[SYN, STY, SEM]$, where the components of the vector are defined

The SYN component constitutes the syntax of the characters. Characters may be either stem characters, as in 'p', or complex characters consisting of a simple character with one or more diacritics, such as 'p^{h}'. The stem character may be analysed in terms of component functions such as circles, descenders and ascenders. The IPA stem characters are represented by a standard coding known as the IPA coding [International Phonetic Association, 1999], sometimes as the *Esling codes* [Esling and Gaylord, 1993], in which each character or diacritic has a numerical code, and the syntax of diacritic arrangements over, under, left and right of characters is defined. Unlike the Unicode code–blocks, the IPA numbers cover the entire IPA character set, and the mappings to IPA semantics and glyphs are technically complete and sound. The IPA code numbers are therefore suitable as a representation at the generic level which we introduce in the present contribution, and for practical purposes these numbers can be mapped into other less straightforward codes (e.g. Unicode, LaTeX macros, TrueType or OpenType font tables).

The STY component constitutes the *style semantics* (rendering semantics) of the character, i.e. a mapping of the character (represented by its Esling code, or its code in another code table) to a glyph (or a glyph structure consisting of an arrangement of glyphs) in the sense already defined. A standard description of the IPA glyphs is provided by Pullum and Ladusaw [Pullum et al. 1986]; this description pre–dates the most recent revisions of the IPA in 1993 and 1996, however. The style semantics is thus an interpretation function from the character syntax into the style semantic domain of glyph configurations: $R : SYN \rightarrow STY$.

The SEM domain constitutes the *domain semantics* of the character, e.g. the sound type denoted by an IPA character as defined by the International Phonetic Association. In the ASCII code set, the hex code 07 denotes a warning, and is rendered by the acoustic beep. The hex code 58 denotes the upper case version of the 24th letter of the English alphabet and is rendered by 'X'. The denotational semantics is thus an interpretation function from the character syntax into the user–oriented semantic domain: $D : SYN \rightarrow SEM$. Examples of denotations of IPA characters are:

- the *voiceless velar fricative* denoted by the simple character 'x',

- the *aspirated voiceless bilabial plosive* denoted by 'p^h'.

In fact, phoneticians define a number of subdomains for the IPA characters, one of which is language independent (the narrow phonetic domain of physical sounds), the others being language dependent (the phoneme sets of individual languages). The narrow phonetic domain is indicated by square bracket quotes [p], and the phonemic domains are indicated by forward slash quotes /p/. The quotes represent semantic interpretation functions from the character rendered by the glyph or glyphs which they enclose into the relevant denotation domain of the character. The mappings $I : SYN \rightarrow SEM$ and $I : SYN \rightarrow STY$ are traditionally defined implicitly and simultaneously in the IPA chart.[1]

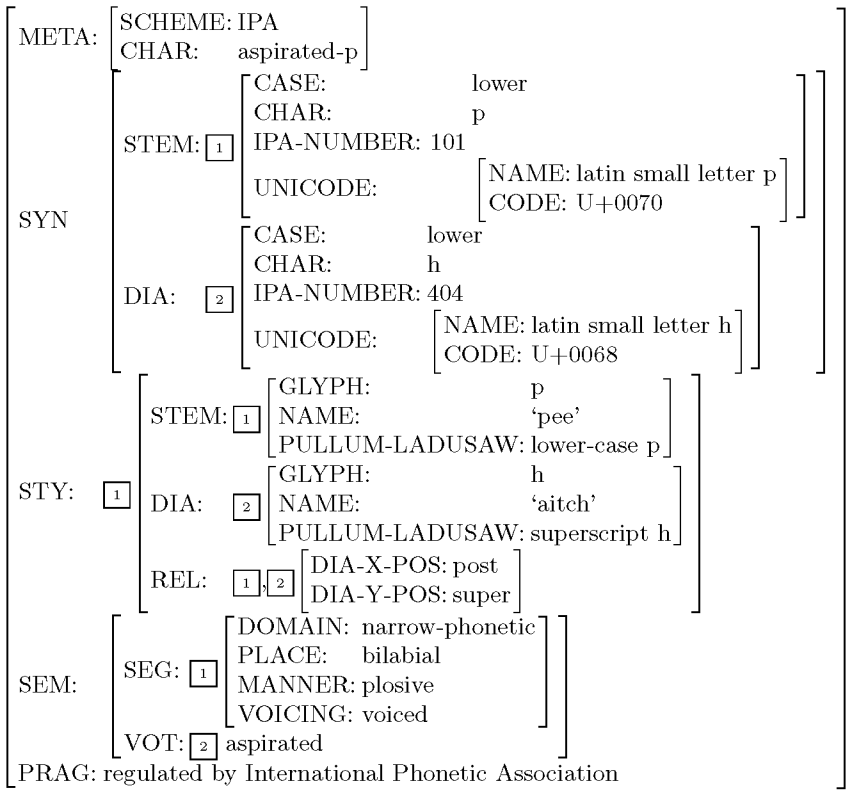

Fig. 1. Structure of semiotic vector extract for [p^h] in IPA name-space.

For IPA characters, the vector components SYN, STY and SEM are further analysed into component–specific vectors specifying syntactic composition, glyph structure style, and sound type semantics respectively. These

[1] `http://www2.arts.gla.ac.uk/IPA/fullchart.html`

vectors can be represented as attribute value structures in a standard linguistic notation; the example illustrated is 'p^{h}'. The composition of the syntactic composition sub–vector SYN, the glyph structure style vector STY, and the sound type semantic vector SEM for 'p^{h}' are shown in Figure 1.

The full description cannot be given in this context for reasons of space. Indeed there may not be a full description, in the sense that alternative codings for this character also exist, in the form of the Esling codes, LaTeX macros, as code–points in legacy fonts, or even as the SAMPA mapping to basic latin characters [Gibbon et al. 2000], and can be included in the attribute–value structure. Following computational linguistic conventions the mappings between the main vectors are shown here by co–indexing the related properties in the three main vectors. The detailed technical formalisation of the mappings between syntax, rendering and semantics is not the subject of this contribution, however.

3 Knowledge Discovery from Character Encodings

Having laid the foundation of characters as complex constituents, described the relationships between characters and higher level constructs such as fonts and explored the various types of properties applied to characters, we can now turn to a discussion of how these properties can be manipulated and explored in various different ways to realise new linguistic knowledge from the underlying characters themselves.

With this analytical and representational mechanism we are able to classify characters from a number of perspectives, including their proximity in the semiotic vector space, in linguistic meaning, structure and context–sensitive rendering, provenance throughout a family of related fonts etc. The details of the nomenclature will no doubt lead to controversial debate, but the architecture of our approach to generic character classification is clear.

From the semiotic vector model illustrated in Figure 1 we can derive a number of different types of classification and relation mining strategies for different application domains:

- multi–dimensional classifications based on similarity of any combination of components of the semiotic vectors;
- computation of tree representations, graph representations or matrix representations for visualisation, search, sorting and merging with standard unification grammar operations;
- similarity definition, 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.

4 Towards a Character Mapping Ontology

The AVM–based metrics can be displayed in a number of ways for interpretation. For mappings to specific fonts we favour an ontological approach, considering character encodings used within a single font as a type of name–space, thus enabling mappings to many different encodings.

In the simplest case, we could utilise the simple character mapping ontology discussed in [Gibbon et al, 2004], which defined an XML data structure for a given character set, and hence the basis on which different character sets could be compared. More complex comparisons and mappings may be expressed in a character markup mapping language eg CMML [Davis and Scherer, 2004].

A fully expressed out character ontology based on the principles outlined in the present discussion requires extensive further discussion in order to achieve a working consensus. As a minimal requirement, a distinction between the SYN, SEM and STY attributes is required; further distinctions, as in Figure 1, will have variable granularity and be extensible on demand.

Assuming coherent definitions of characters as signs with SYN and SEM attributes for a particular character set (of which the IPA code numbers and their definitions as given in [International Phonetic Alphabet, 1999] are a suitable example), the remaining issue is how to map the syntactically and semantically coherent system into other encodings, both into Unicode and into code points for glyph collections in specific fonts. At the present state of the art, there are two options, such as the following for the IPA:

1. Mapping of IPA code numbers directly into code points (or sets of code points) in specific fonts such as IPAKIEL, SILDOULOS or TIPA.
2. Mapping of IPA code numbers into Unicode, in which codes may be scattered over different code–blocks, with a second layer of mapping into specific fonts.

If these mappings are known, then in principle the properties defined in the ontology can be associated with other encodings and their glyph renderings. But note that with the current Unicode regime, an inverse function is not available: since the basic latin codes are massively ambiguous with regard to their SEM, i.e. user–oriented semantic, properties, there is no simple way of inducing a mapping from glyphs, or even from Unicode numbers, into the semantically oriented encoding. In this respect, Unicode numbers are no different from the codes for glyphs in any arbitrary font.

The solution to this problem is to map ontological codes to font code–points with a convention such as name–space assignment. A biunique mapping is created by distinguishing between, say, 'ipa:basic_latin' (the IPA relevant subset of the basic latin code block) and 'english_alphabet:basic_latin' (the subset containing the 52 upper and lower case characters of the English alphabet) or 'ascii_keyboard:basic_latin' (the subset including digits, some punctuation marks and some cursor control codes). The IPA Unicode mapping is then from the ontological representation to the union of two character blocks: $ipa : basic\_latin \cup ipa : ipa\_extensions$.

5 Future Directions

For the purpose of defining interoperable text processes over characters, these mappings can be expressed straightforwardly in XML and manipulated at the levels of ontology, unicode, font and glyph properties by an appropriate language such as XSL. The next steps in the present enterprise are:

1. translation of the formal properties illustrated by our IPA example into interoperable XML;
2. definition of inter–level mappings between ontological information and both Unicode blocks and specific fonts;
3. development of an encoding definition language as a tool for specifying the $< SYN, STY, SEM >$ vector and its subvectors;
4. practical characterisation of the properties of legacy documents which use non–standard fonts.

6 Conclusion

The analytical and representational model presented here permits complex data mining operations over linguistic data regardless of its expression in particular character encodings. Furthermore, the approach permits complex linguistic properties to be used coherently as query terms, a dimension not associated either with legacy fonts or Unicode.

Using a semiotically based ontological approach to character encoding, a new dimension to the definition of text processes for search and text classification can be defined. For example, an electronic document which contains

uses of a font such as IPAKIEL or SILIPA can be assigned to the semantic domain of linguistics with a high degree of confidence, and can thus be assumed to have been authored by a linguist with that degree of confidence. This is only the case, of course, if the relation between the font and the relevant ontology has been defined. The same applies to other specialised fonts which relate to other semantic domains, with far–reaching consequences for document classification in the context of the semantic web.

With an ontological approach to character description of the kind introduced in the present contribution, generic search tools can be developed with a far higher degree of granularity than is currently available. An important issue for future work will be how the development of ontologies of this kind can be supported by machine learning techniques. Given that characters are the smallest units of text, they are available in sufficient numbers to permit the application of sophisticated induction techniques for this purpose.

References

DAVIS, M and SCHERER, M. (2004): Character Mapping Markup Language (CharMapML). Unicode Technical Report #22, Unicode Consortium. `http://www.unicode.org/reports/tr22/`

DÜRST, M., YERGEAU, F., ISHIDA, R., WOLF, M. and TEXIN, T. (2005): Character M odel for the World Wide Web 1.0: Fundamentals. World Wide Web Consortium. `http://www.w3.org/TR/charmod/`

ESLING, J. H. and GAYLORD, H. 1993. Computer Codes for Phonetic Symbols. *Journal of the International Phonetic Association* 23(2), pp. 83–97.

GIBBON, D., BOW, C., BIRD, S. and HUGHES, B. (2004): Securing Interpretability: The Case of Ega Language Documentation. *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, 2004.* Euopean Language Resources Association: Paris. pp 1369–1372.

GIBBON, D., MERTINS, I., MOORE, R. (2000): Handbook of Multimodal and Spoken Language Systems: Resources, Terminology and Product Evaluation. New York etc.: Kluwer Academic Publishers.

HIMMELMANN, N. P. (1998): Documentary and descriptive linguistics. *Linguistics* 36, pp.161–195.

INTERNATIONAL PHONETIC ASSOCATION (1999): Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press: Cambridge. `http://www2.gla.ac.uk/IPA/`

PULLUM, G. K. and LADUSAW, W. A. (1986): Phonetic Symbol Guide. The University of Chicago Press: Chicago.

UNICODE CONSORTIUM, (2003): The Unicode Standard, Version 4.0, Reading, MA, Addison–Wesley, 2003. `http://www.unicode.org/versions/Unicode4.0.0/`