ORTHOGRAPHIC VARIATION IN LEXICAL DATABASES

Maarten Janssen

Rua Conde de Redondo 74 – 5º
1150-109 Lisboa, Portugal

ABSTRACT

Traditionally, orthographic variants have been modelled as different ways of spelling the same word – described at the level of the lexeme. But when inflection is taken into account, this runs into a problem: different citation forms have different inflectional paradigm – and orthographic variation does not merely affect the citation form, but the entire paradigm. The MorDebe database therefore models orthographic variation as a relation between distinct, yet still token-identical lexemes. This paper discusses the advantage of that approach, and the full set of practical problems that arose during the structural treatment of orthographic variation in the MorDebe database.

Orthographic Variation in Lexical Databases

1 Introduction

Dictionaries are most commonly used simply to see if a word exists, and how it should be written – not to look up its meaning. It is used for this purpose in some 70% of the cases according to Oppentocht & Schutz (2004). So dictionaries do have a normative force with respect to spelling, since users tend to follow the spelling presented in the dictionary. And in some cases the dictionary even has a real normative character – as as is the case with official guides like the *Woordenlijst der Nederlandse Taal* (GB). De/Prescribing the correct spelling, dictionaries have cleared up much of the orthographic multiplicity that existed some centuries years ago.

But in many cases, this prescriptive character is not completely univocal – for some words there are alternative ways of writing the same word both of which are accepted by the dictionary – the GB of 1990 allows the Dutch word for gift to be written either as *cadeau* or as *kado*, and most English dictionaries allow both *medieval* and *mediaeval* as correct spellings. This is what we refer to in this article as *orthographic variation*: alternative correct ways of writing the same word.

An important aspect of orthographic variation is that they concern the same word – the identity of the lexeme is not affected by its orthographic realization. Orthographic variation is not a strong type of synonymy where there are two words that fully express the same thing – the two words involved in orthographic variation are token identical, not type identical as in the case of synonymy. For paper dictionaries, this is not a highly relevant point: when the two variants are not alphabetically next to each other, one of the lexical entries is listed as a cross reference to the other – but this is the same way in which dictionaries often treat strong synonymy, and irregular inflection. Cross-referencing in dictionaries is a way of facilitating finding words, as well as a way of saving space, and can be caused by a variety of factors. So from the perspective of traditional dictionaries, the exact status of orthographic variants is not very relevant.

But for the design of lexical databases the issue of token-identity of orthographic variants is more crucial – especially in the case of highly structured lexical databases. If the database makes a distinction between a lexemes and meaning, then in the case of synonymy, the things that should be related are the meanings, whereas in the case of orthographic variation, the cross-reference should be made at the level of the lexeme itself – the two lexemes have to be treated as one.

This paper will focus on the problem of the interaction between inflection and orthographic variation, and discuss some smaller problems of a more practical nature, and their solution within the MorDebe database. Most of the examples in this paper are also taken from the MorDebe database, and are hence Portuguese examples.

2 Inflection and Orthographic Variation

Since orthographic realization does not alter the identity of the word, variation should be seen as a phenomenon *within* the "word". We say that the same word *medieval* can also be written as *mediaeval*. This is why most lexical databases treat variation within the lexical entry – the Göteborg Lexical Database lists spelling variants as part of the formal characterisation of the lemma, the Pronunciation Lexicon Specification (Baggi, *unpublished*) a single lexeme can have more than one grapheme. The matter of orthographic variation gets very little attention, since it is seen as non-problematic. However, this is only true as long as inflection is not taken into account.

Since orthographic variation relates to orthography, it is strongly related to word-forms, since word-forms are the entities that are written down, not the lemmas. It is the word-forms that can be written in different ways in some cases - in Portuguese, the 1st person singular present tense indicative of *ouvir* (hear) can be written either as *ouço* or as *oiço*. Both forms are independently a correct form of *ouvir*. From the perspective of representation, there is very little that distinguishes this othographic variation from other types of inflectional variation – such as the fact that the past tense of the Dutch verb *waaien* (to blow) can either be produced regularly (*waaide*), or irregularly (*woei*). There are simply two ways of writing the same form.

In the same manner, we could say that the word-forms *kados* and *cadeaux* are two different realisations of the same inflectional form of the lexeme *kado*, where there are also two form of the singular: *kado* and *cadeau*. But in the case of *kado* there is a problem: since all four forms are independently the plurals and singulars of a single lemma, they are all linked in the same fashion.

And this would come down to saying that *kado* and *cadeaux* are the singular and plural form of each other.

In a sense, we could say that *cadeaux* is the plural of *kado* – in the sense that *cadeaux* is a possible form of the plural, the corresponding singular of which has a form *kado*. But in that case, we should say that *kado/cadeaux* is the plural of the singular *kado/cadeau.* And that in turn would imply that the inflected form itself is an abstract entity, which can have various orthographic realisations. To get to the actual written form, we would need to stipulate an additional level below the inflected forms. Apart from this, this would create the need to say that the citation form of this dual orthography paradigm is *kado/cadeau* – which would at best complicate the status of citation forms.

A different solution would be to say that there is a single lexeme *kado* has two associated inflectional paradigms – the citation forms of which are resp. *kado* and *cadeau*. But in that solution, we would also have to adopt a 3-layered structure: on the top, there is the lexeme, below which are the inflectional paradigms, represented by their citation form, and below that are the actual word-forms. In order to avoid this, orthographic variations in MorDebe are in fact modeled as separate lexeme, with a strong relation between them.

2.1 Orthographic Variation in MorDebe

MorDebe is a large-scale database of lexical information – with an emphasis on inflectional morphology. The database basically consists of a two-table structure – one table containing the lemmas with their grammatical class and citation form, the other table containing the inflected forms with their orthography and their inflectional form. Every word-form is linked to a single lemma, and the full set of inflectional forms related to a lemmas providing its inflectional paradigm.

The way variation is treated within this structure is as follows: when there is orthographic variation within an inflectional paradigm, there are two competing inflectional forms, both linked with to the same lemma with the same inflectional coding. As an example: both *oiço* and *ouço* occur in the table of inflectional forms, and both are linked to the lemma *ouvir* as its *presInd1s* – first person singular present indicative. But when the variation is at the level of the lemma, there are two distinct lemmas in the main database, each of which has a full inflectional paradigm.

In the latter case, the fact that the two lemmas are in fact variation of the same lexeme, the two entries in the database are linked by means of a relation *alt*, which explicitly models the fact

that these two lemmas are in fact token-identical. The relation between the two orthographic variants is assymetric: one of the entries will function as the main entry, the other as a variant. The status as main entry can be due to the fact that it is the prefered spelling (as in the case of *cadeau*), or in the absence of a prefered spelling it is the most prominent in terms of frequency. The secondary entries do not need to be adorned with grammatical or semantic information since they inherit these from the main entry.

In some cases, there is more than one secondary form, as in the extreme case of the Portuguese *luzencu* (glow-worm, *regional*), which can also be written as *luzincu, luzecu, luzicu,* or *luze-cu*. Because of the assymetry of the *alt* relation, we do not need the full set of 10 possible relations between these forms, but all secundary spellings are only linked to the main entry. This is very similar to the solution adopted in many dictionaries, where all secondary entries refer to a single main entry.

The fact that the relation by which orthographic variants are linked in MorDebe (alt) models token-identity is not an inherent feature of the relation, but an interpretational matter. In fact, the same structure is used to model inherent inflections: in Portuguese, many grammar books describe the variation between the male and female forms of nouns (such as *gato* – cat/tomcat and *gata* – she-cat) is seen as inflectional rather than derivational. But other sources quote it as a derivational relation. Also in these cases, the male and female inflectional paradigms are listed as separated entries, linked by a relation *fem*. And the same is done in other cases that are somehow between derivation and inflection such as *aum* for aumentative forms, *dim* for diminutive forms, *s0* for deverbal nouns, *s0a* for deadjectival nouns, etc. (Janssen, 2005).

3. Practical Problems

This section will discuss some problems of a more practical nature, which are the kind of problems that have to be dealt with when taking orthographic variation serious in the design of a lexical database. Along with these problems it will present the way these problems are dealt with in the MorDebe database.

3.1 Variation and Normativity

As said in the introduction, dictionaries are up to a large degree normative with regard to orthography. Although the dictionary will often only try to reflect the correct spelling of the language, the general public uses the dictionary to verify how words should be written, hence assigning a normative role to the dictionary. Therefore, dictionaries should not contain words that are considered badly spelled.

But although MorDebe does pretend to be useable as an orthographic guide, it is by design intended to be a multi-purpose database. And leaving out bad spellings is not a good solution from every perspective. To take a concrete example, the fact that in Dutch the word *gazelle eye* had to be written like *gazelleoog* until 1995, after that as *gazellenoog*, but since 15-10-2005 again as *gazelleoog* (according to the changing official spelling in GB) changes the way the word will have to be written in new official texts, but it does of course not change existing texts. Similarly, you will find the word *kado* in Dutch texts (both old and new) even though it is no longer considered a correct spelling. Therefore, for a user looking for the meaning of a word he encountered, it would be helpful to be able to find these words in the dictionary even though they are not considered correct (anymore).

So MorDebe faces two incompatible desires: on the one hand the need to incorporate only correctly spelled words in order to be able to inform the users about the correct spelling of words – and on the other hand the desire to inform the user about any word he encounters in a text independently whether it is considered (fully) correctly spelled at that time. To solve this problem, MorDebe introduces the notion of *graded variation*. When linking a word as an orthographic variant of another, it is classified either as a fully equivalent variant, or as a non-preferred variant, or even a not or no longer correct variant. This solution is similar to the way infrequent or non-prefered words are listed in for instance the Houiass (2003) dictionary. In the interface, the correct variants will be indicated in both direction, but non-preferred variants will only refer to their preferred spelling, not the other way around.

3.2 Regular Variation and Dialect

Portuguese, like English, is a language with a dual orthographic standard: the orthographic rules in Brasil differ slightly from those used in Portugal, Timor, and the African countries (Angola, Moçambique, São Tomé, Cabo Verde). There are four major different between Brasilian Portuguese (PB) and what is most often called European Portuguese (PE): firstly, PB graphically marks the

difference between a *u* following a *q* or *g* where the *u* is pronounced as a [w] and where it is not pronounced by mean of a diaeresis on the *u* in the former case. So PB writes *delinqüir*, where PE has *delinquir*. Secondly, PB drops the *c*, *p,* and *m* in front of a *t, c,* or *n* where the *c/t/m* is not pronounced: *elétrico* vs. *eléctrico*, *ação* vs. *acção*. Thirdly, PB graphically indicates the nasal *o* or *e* in front of an *m* or *n*: *acrômico* vs. *acrómico*. And finally, PB has a stress mark in tritongues *eio* and *aio*: *açotéia* vs *açoteia*.

The problem with these dialectic variations is that they do not present cases in which both orthographic realisations are correct in identical circumstances, but where the correct orthography depends on the dialect used: one will only be correct in a text written in PB, the other only in PE. For this reason, these dialectically dependent orthographic variations are modeled using a distinct relation. This allows providing the correct information to the user: if the user is looking for PE spelling, the interface will refer from the PB word to its correct PE spelling – mentioning that this is the PB variation of the word. But in the other direction, the interface will not refer from the PE spelling to the (incorrect from the user perspective) PB spelling.

3.3 Meaning-dependent variation

A good argument in favour of the view that orthographic variation is a lexeme-based phenomenon and not a string-based phenomenon is the fact that orthographic variation can be meaning dependent: the Portuguese word *camareiro* is an orthographic variation of the word *camaroeiro* (shrimp fisher) – which a derived from the word *camarão* (shrimp). But there is a homonym *camareiro* (chamberlain), derived from *câmara* (room), which does not display this variation. The noun *loura* (blond woman) can also be written as *loira*, but only the homonym *loura* (burrow) can be written as *lura* as well.

The cases above are all clear cases of homonymy – but there are also cases in which the orthographic variation only appears with certain meanings of a word – even though such cases are always marginal. The word *leader* as a loanword in Portuguese is written only as *líder* these days – even though dictionaries still accept *leader*. But the word *leader* is basically only used as a technical term. The word *carácter* in its meaning of the character of a person can only be used in that form, but although not accepted (yet) by the dictionary, it starts being used as *caracter* when a printed character is meant. The variation *cousa* for *coisa* (thing) is slightly archaic – and in its popular use for sexual organ it will never be used as *cousa*.

4. Conclusion

In this paper I hope to have shown that modelling orthographic variation of words not in the traditional way as a lexeme-internal phenomenon, but as a relation between different lexical entries with distinct inflectional paradigms is the most coherent and intuitively plausible way of treating variation in lexical databases. This not only from a theoretical perspective, but form the perspective of its application in a large-scale lexical database project – MorDebe.

Not treated in this paper are borderline cases between (strong) synonymy and orthographic variant such as the pair *couro/coiro* (leather – changing only spelling but also pronunciation), the shortened pairs such as *agar/agar-agar* (a kind of seaweed) and *estar/tar* (to be) or heavily modified variants such as *aguado/ougado* (watered down).

There are additional problems related with orthographic variation in a fully language-independent perspective. An issue that has not been mentioned in this paper is the regular alternation in languages with alternative writing systems, such as the fact that in Japanese all Kanji characters can also been written out in Hiragana, or the fact that Bosnian, Croatian, and Serbian can be found written in either the Latin alphabet, Cyrillic, or in Bosnia even in Arabic. Such completely productive variation should probably be modelled in a dedicated manner. But for single alphabet languages like Portuguese, the method described in this paper is fully functional for the large-scale MorDebe database.

References

A. Dictionaries

Casteleiro, J. M. (2001). *Dicionário da Língua Portuguesa Comtemporânea.* Lisboa, Verbo. (DLPC)
Teixeira, G. (2004). *Grande Dicionário da Língua Portuguesa*. Lisboa, Porto Editora. (GDLP)
Villar, Mauro de Salles (2001). *Dicionário Houaiss da Língua Portuguesa*. Lisboa: Temas & Debates. (Houiass)
Instituut voor Nederlandse Lexicologie. (1995) *Woordenlijst Nederlandse Taal.* Antwerp, Standaard Uitgeverij. (GB)

B. Other Literature

Baggia, P. (unpublished). *Pronunciation Lexicon Specification (PLS)*. W3C Working Draft, feb. 2005.
Janssen, M. (2005) 'Betweeen Inflection and Derivation: paradigmatic lexical functions in morphological databases'. In Apresjan, Ju. D. Iomdin L. L. (eds.) *East West Encounter: Second International Conference on Meaning ⇔ Text Theory.* Moscow: Slavic Culture Languages Publishing House. 187-196.
Oppentocht, Lineke & Rik Schutz. 2003. "Developments in Electronic Dictionary Design". *In:* P. van Sterkenburg (ed.) *A Practical Guide to Lexicography.* Amsterdam: John Benjamins Publishing. 215-227.