How to build a multilingual inheritance-based lexicon

Carole Tiberius

Surrey Morphology Group
Linguistic and International Studies
University of Surrey
Guildford, Surrey, UK, GU2 7XH
c.tiberius@surrey.ac.uk

Abstract

This paper discusses a fairly new approach to multilingual lexical representation which abstracts away from the traditional MT architecture to multilingual lexicons. Rather than linking the monolingual lexicons at the level of semantics only, we aim to construct a multilingual lexicon in which information can be shared at all levels of linguistic description using an inheritance-based formalism. In this paper, we present different architectures that can be used to build such a multilingual inheritance lexicon. Two main approaches are contrasted, a non-parameterised and a parameterised. In a non-parameterised approach, language is not used as a parameter in the multilingual lexicon. The multilingual lexicon consists of a set of monolingual hierarchical lexicons plus a shared hierarchical lexicon containing what the monolingual lexicons have in common. In a parameterised model, on the other hand, all information is integrated into one single hierarchy, and language is used as a parameter to indicate which parts of the hierarchy are valid for which language. The advantages and disadvantages of the different approaches will be discussed with respect to small sample fragments of Dutch, English, Danish, and Icelandic nouns implemented in DATR (Evans and Gazdar, 1996).

1. Introduction

So far most of the work on the application of inheritance networks to multilingual lexical description has concentrated on sense linkage between essentially monolingual lexicons (Copestake et al., 1992), similar to the work that has been done on multilingual lexicon development for practical applications such as MT and multilingual Natural Language Generation. Little attention has been paid to the use of inheritance networks to share information between languages at levels of linguistic description other than semantics.

However, it is well-known that languages (especially related languages) also possess similarities in their syntax, morphology, phonology, etc. An example of syntactic similarities can be found at the level of subcategorisation frames, which often exhibit identical argument slots and similar, if not identical, argument types. Compare the subcategorisation frames of the verb 'to see' in Dutch and English (Kruger and Heid, 1996, p.12) :

[PERCEIVER non-intentionally] see [actual entity PERCEIVED]
English He saw tears in her eyes
Dutch Hij ziet tranen in haar ogen

Morphological similarities between Dutch, English, and German can be found in the declension of a set of subregular verbs in these languages. Compare the forms of the verb *sing*:

English	s**i**ng	s**a**ng	s**u**ng
Dutch	z**i**ng	z**o**ng	gez**o**ngen
German	s**i**ng	s**a**ng	ges**u**ngen

In all three languages there is a change of the vowel from the present to the past tense and in English and German this vowel is even the same. Many more similarities can be found and all these similarities could be captured in a multilingual inheritance network. As argued by Cahill and Gazdar (1999), capturing such similarities could contribute significantly to the robustness, maintainability, and extensibility of multilingual NLP systems.

The idea of capturing similarities at different levels of linguistic description in an inheritance-based framework has previously been explored in Kameyama's (1988) multilingual unification grammar, in the PolyLex project (Cahill and Gazdar, 1999), and in the GREG project (Kilgarriff et al., 1999). All these projects use what we call a non-parameterised approach. In this paper we will contrast this with a parameterised approach to multilingual lexical representation.

The remainder of this paper is organised as follows. Section 2 defines different architectures that can be used to build a multilingual inheritance lexicon following proposals of Evans (1996). These architectures will then be evaluated with respect to small sample fragments that have been implemented in DATR in Section 3. In Section 4 we discuss the implications of our results and in Section 5 we give conclusions.

2. The Multilingual Architectures

In this section, we discuss various architectures that can be used to construct a multilingual inheritance-based lexicon. First, we show how inheritance techniques that are generally used in a monolingual context can be extended to the multilingual case. For a general introduction to inheritance-based formalisms, the reader is referred to Daelemans and Gazdar (1992). In our multilingual lexicons we assume an orthogonal non-monotonic multiple inheritance network as is illustrated in Figure 1[1]. That is a node in the hierarchy can inherit information from more than one parent node as long as this information is distinct (e.g. *hate* inherits its syntactic properties from the

[1]This figure is based on example networks given in Daelemans, De Smedt and Gazdar (1992).

TRANSITIVE_VERB class and its morphological properties from MOR_VERB) and information can be overridden lower down in the hierarchy (e.g. the past participle information which is specified at the Beat node overrides the past participle information inherited from the top of the hierarchy, resulting in a past participle *beaten* rather than *beated*.).

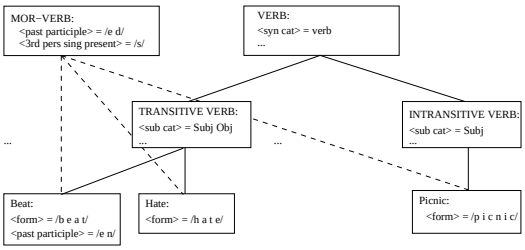


Figure 1: Non-Monotonic Multiple Inheritance Network

In a multilingual inheritance network we do not only need to be able to capture generalisations within languages but also across languages. In order to capture cross-linguistic generalisations, the inheritance network needs to have some kind of means to indicate that parts of the network are valid for more than one language. That is, parts of the lexicon need to be linked to a language typology which can be more or less complex on the basis of whether languages are grouped together into classes or not, for example based on genetic relationships that exist between languages. Such a language typology can be modelled using the same techniques as in monolingual inheritance networks. For example, a lexicon for Dutch, English, French, and Spanish might use the following language typology:

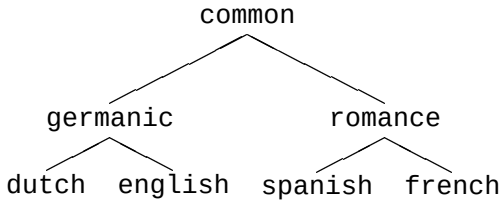


Figure 2: A language typology

Information which is shared by all those languages will be associated with common. Information which is specific to the Romance languages will be associated with romance and information which is specific to French or Spanish will be associated with respectively french or spanish. In this multilingual lexicon, it might be reasonable to suppose that french and spanish inherit from romance, dutch and english inherit from germanic, and both germanic and romance inherit from common. These inheritance relations can be monotonic or non-monotonic. In our lexicons, we assume that they are non-monotonic. Thus, information inherited from, for instance, Germanic can be overridden for English, etc.

We now turn to how parts of the lexicon can be linked to such a language typology. Evans (1996) distinguishes two approaches which he calls parameterised and non-parameterised.

2.1. The Non-Parameterised Model

In a **non-parameterised** model, the multilingual lexicon is constructed by taking a set of monolingual hierarchical lexicons and creating a parallel hierarchy containing what the monolingual lexicons have in common. The resulting structure for a multilingual lexicon with a flat language typology is illustrated in Figure 3. In this figure, we

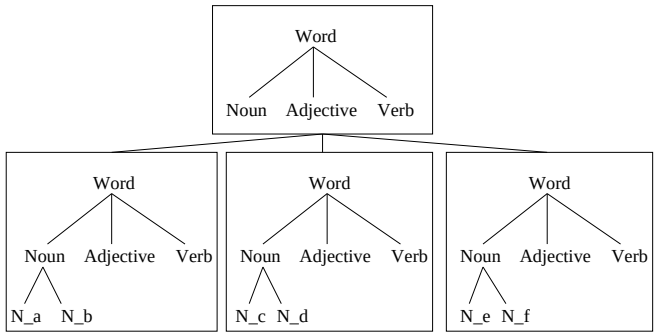


Figure 3: Non-parameterised multilingual inheritance hierarchy with a flat language typology

have the hierarchical lexicons of three different languages at the bottom of the figure and a shared hierarchy at the top. We see that all three monolingual hierarchies have a Word, Noun, Adjective, and a Verb class. This shared information is captured in the shared hierarchy at the top of the inheritance network. The structure of a multilingual lexicon where languages are grouped into classes is given in Figure 4. In this figure, two of the three languages share

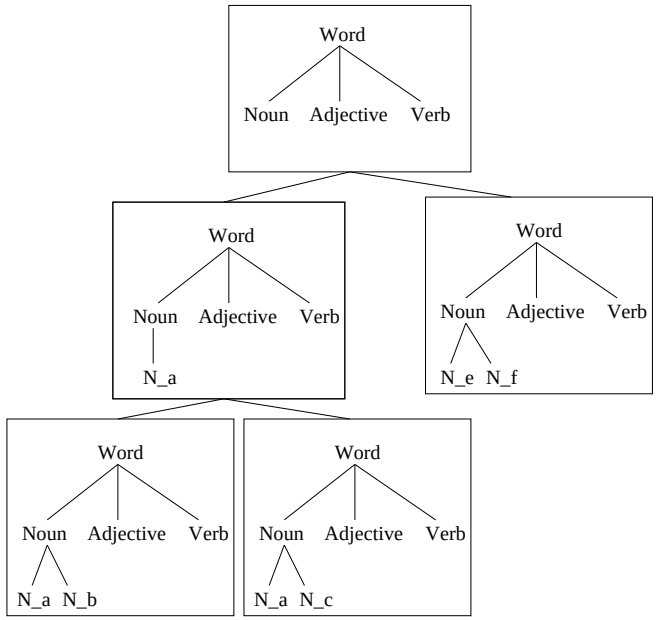


Figure 4: Non-parameterised multilingual inheritance hierarchy with subhierarchies

the N_a subclass. This generalisation is captured by group-

ing these two languages together into a subhierarchy. Such a subhierarchy could correspond to a (sub)family of languages. Evans calls these networks non-parameterised because language is not explicitly used as a parameter. There is in principle nothing which ties a particular hierarchy to a particular language in the multilingual inheritance structure. Each hierarchy belongs to an individual language or represents information shared by a set of languages, but nothing in the hierarchy tells you explicitly which language or languages are concerned – the knowledge of the different languages involved is in the user's head rather than in the theory.

Evans also calls the non-parameterised model the Structure-Sharing model. We will use this term in the remainder of this paper. The Structure-Sharing model is essentially the model that has been used in the PolyLex (Cahill and Gazdar, 1999) and GREG (Kilgarriff et al., 1999) projects.

2.2. The Parameterised Model

In a parameterised model, on the other hand, all the languages represented in the lexicon are integrated into a single hierarchy and language is used as a parameter to indicate which parts of the lexicon are valid for which languages. A schematic illustration of a parameterised model is shown in Figure 5. The different boxes indicate which

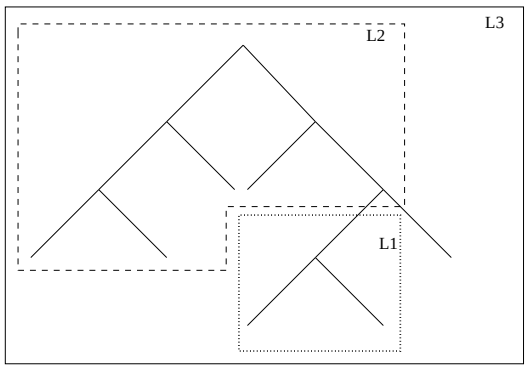


Figure 5: Parameterised multilingual inheritance hierarchy

part of the hierarchy is valid for which language. The whole hierarchy is valid for language `L3`, the dashed line indicates the part which is valid for language `L2`, and the dotted line indicates the part that is valid for language `L1`.

Following Evans' proposals we focus on parameterised models in which language parameters are inserted in the feature theory. For the purposes of the present discussion, we assume that the language parameters are organised in a tree structure as represented above for Dutch, English, French, and Spanish.

Thus a parameterised model consists of a feature tree for a particular structure (e.g. noun features as illustrated below in Figures 7 and 8) and a language tree representing the language typology. The question that arises is how can these two trees be combined? In other words, where in the feature tree can language parameters be inserted and once language parameters have been inserted, how can inheritance be made to work in both the language tree and the feature tree. Evans suggests three models, which we will discuss below. First, the language tree can be inserted at the bottom of the feature tree, the Micro-Features model. Second, the language tree can be inserted at the top of the tree, the Meta-Features model. Finally, he hints at a third model in which the language tree can be inserted at any point in the feature tree. He calls this the Infinitesimal model.

2.2.1. The Micro-Features Model

In the Micro-Features model, language parameters occur at the bottom of the tree as is illustrated in Figure 6. Generally in inheritance networks, the lower the position

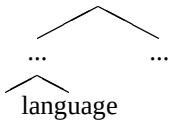


Figure 6: Illustration of the Micro-Features model

in the hierarchy at which a property appears, the more exceptional it may be considered. Thus, the Micro-Features model is based on the assumption that variation between languages is exceptional rather than rule. It assumes one shared feature tree for all languages with only local, low-level variation occurring at the bottom of the tree. However, generally, different languages have different feature trees and there are higher level language-dependent generalisations, even between closely related languages, that an adequate multilingual lexicon should be able to capture. This is illustrated below with the feature trees for the noun features in Dutch, English, and German. Nouns only inflect for `number` in Dutch and English, whereas they inflect for `number` and `case` in German.

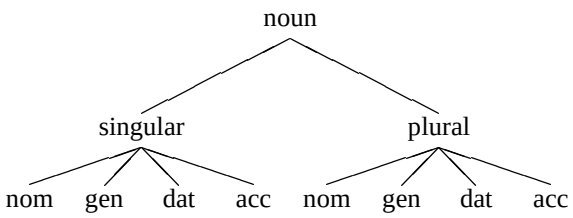


Figure 7: Feature tree for nouns in German

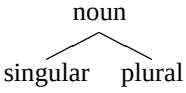


Figure 8: Feature tree for nouns in English and Dutch

The Micro-Features model cannot deal with this situation. For the Micro-Features model, the feature tree has to be the same (i.e. have the same features, not necessarily the same feature-values) up to the point where language is inserted, which is completely at the bottom in the Micro-Features model. Thus the Micro-Features model cannot capture higher level generalisations such as that singular nouns in Dutch and German are subject to final devoicing

whereas they are not in English. The applicability of the Micro-Features model is therefore limited and it will not be further considered as a viable option for constructing multilingual inheritance-based lexicons.

2.2.2. The Meta-Features Model

The Meta-Features model does the opposite of the Micro-Features model and language parameters occur at the top of the tree as is shown in Figure 9. Thus, the Meta-

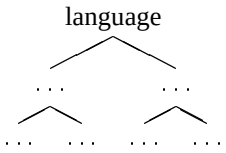
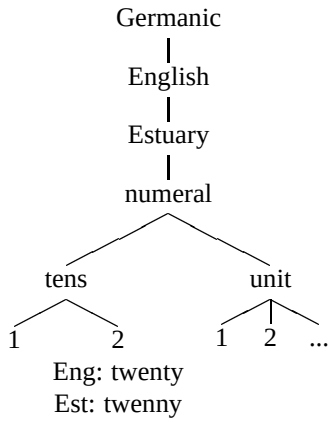


Figure 9: Illustration of the Meta-Features model

Features model is good at capturing higher level language-dependent generalisations such as nouns in one particular language have a property `x`, whereas nouns in general have a property `y`.

As inheritance relations exist in both the language tree and the feature tree, the Meta-Features model is also good for expressing minor variations between languages. For example, adding a new dialect to the lexicon which is related to one of the languages already encoded in the lexicon, requires a change in the language typology, but it does not necessarily affect the feature tree. This is illustrated below with an extract of a feature tree for numerals in English and Estuary, a dialect of English (Evans, 1996).



Here, English and Estuary have the same feature tree with different values for <`numeral tens 2`>. Estuary will inherit all information that is specified for Germanic English, except for the value of <`numeral tens 2`> which is *twenny*.

Thus, in the Meta-Features model, the feature tree can either be completely the same or completely separate. Generalisations at intermediate levels cannot be captured. This means that the Meta-Features model does not allow us to capture the fact that Dutch and German singular nouns are subject to final devoicing whereas English nouns are not either.

2.2.3. The Infinitesimal Model

The Infinitesimal model combines the features of the Micro-Features model and the Meta-Features model. Language parameters can occur at the top, at the bottom of the tree or anywhere in between.

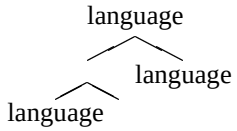


Figure 10: Illustration of the Infinitesimal model

Thus, language-specific characteristics can be captured at any level in the tree, completely at the top (to capture that nouns in language 1 behave differently from nouns in language 2), or completely at the bottom (to capture for example that singular nominative nouns in language 1 do something different from singular nominative nouns in language 2), or anywhere in the middle where one language behaves differently from the other(s). This makes the Infinitesimal model potentially the most powerful model as it allows one to capture language variation anywhere in the hierarchy.

An example of the Infinitesimal model is shown below with a tree structure for the noun features for German and Danish. In Danish, nouns inflect for `number` and `definiteness`, whereas they inflect for `number` and `case` in German. The different feature trees are integrated into

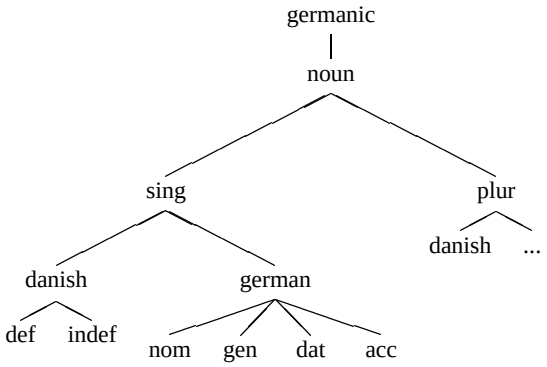


Figure 11: Parameterised tree structure for a subset of the Germanic languages

one shared tree with a part that is specific to Danish and a part that is specific to German.

In the next section, we discuss the advantages and disadvantages of these models by comparing sample implementations covering a small set of nouns in Dutch, English, Danish, and Icelandic.

3. Implementation and Evaluation

Sample lexicons have been implemented and tested running Sussex/Brighton DATR-2.8[2] under Sussex Poplog Prolog[3]. They cover a small set of body part terms in four Germanic languages – Dutch, English, Danish, and Icelandic. The fragments focus on the sharing of morphological, phonological, and morphophonological similarities

[2] See `http://www.datr.org`.

[3] See `http://www.poplog.org`.

between these four languages. Those levels were chosen for illustratory purposes, but the same principles can be applied to other levels of linguistic description. An extract of the dataset is given in Table 1.

Lexeme	**English**	**Dutch**	**Danish**	**Icelandic**
foot	/fUt/ (‘foot’)	/vu:t/ (‘voet’)	/foD/ (‘fod’)	/f@Ut-hYr/ (‘fóstur’)/
mouth	/maUT/ (‘mouth’)	/mOnt/ (‘mond’)	/mon/ (‘mund’)	/mYn/ (‘munni’)
nose	/n@Uz/ (‘nose’)	/n2:s/ (‘neus’)	/nes@/ (‘næse’)	/nE:v/ (‘nef’)
heel	/hi:l/ (‘heel’)	/hi:l/ (‘hiel’)	/hel/ (‘hæl’)	/haId_0l_0/ (‘hæll’)

Table 1: Extract of dataset

In the sample fragments we used the lexical description framework described in Tiberius and Evans (2000). This framework supports the description of lexical generalisations traditionally modelled as morphology and phonology in a single phonological feature based representation. It organises the lexicon into distinct self-contained modules corresponding to levels of lexical description (lexemes, syllable sequences, syllables, and phonemes) as is illustrated in Figure 12 for the Dutch lexeme *Gebed* (‘prayer’). As a lexeme, *Gebed* is primarily linked into the lexeme hierarchy, inheriting from `Noun_EN`, a subclass of `Noun`. But it inherits part of its content, namely its phonological form, from `GEBED` in the syllable sequence hierarchy. `GEBED` is primarily a `Disyllable`, but it inherits part of its content, namely the two syllables it contains, from `GE` and `BED` in the syllable hierarchy. Finally the syllable `BED` inherits part of its structure, from the consonants `b` and `d` and the vowel `E` in the phoneme hierarchy.

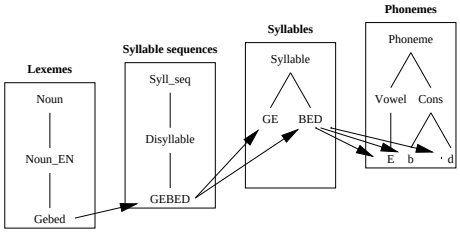


Figure 12: Module and node structure for lexeme *Gebed*

Each of these modules forms its own independent inheritance hierarchy such that generalisations can be captured at each level. Higher level relationships between word forms are represented by means of lexical rules. This way, the framework provides a flexible means of capturing lexical generalisations within and across languages. In the remainder of this section, we first discuss the characteristics of the implementation of the different models and then we turn to their evaluation.

3.1. Implementation

3.1.1. The Structure-Sharing model

The Structure-Sharing fragment was implemented using a non-parallel development strategy. That is, all four monolingual lexicons were first fully developed separately before they were integrated into a multilingual lexicon capturing the similarities that exist between them. An extract of the hierarchical structure of the Structure-Sharing lexicon is given in Figure 13. Note that in our lexicon we use language identifiers to keep track of the different languages in the lexicon, viz. D for Dutch, DK for Danish, E for English, and I for Icelandic. Figure 13 only contains

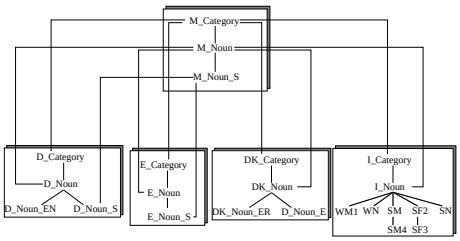


Figure 13: Lexeme hierarchy in the Structure-Sharing lexicon

part of the inheritance hierarchy (similar inheritance hierarchies exist for the syllable sequence, syllable, and phoneme modules), but it is already clear from this picture that there is a lot of redundancy in this network. Each language has its own hierarchy and the inheritance pattern within a language is repeated over and over again.

Our fragment only covers a small set of lexical entries and one can imagine that when the lexicon becomes bigger and more languages are involved, the hierarchical structure and the interactions between the different hierarchies would become even less transparent.

3.1.2. The Meta-Features model

We saw above that in the parameterised models, language features are inserted in the feature theory. Our lexical description framework divides the feature space into two parts: a lexical rule part and an object part. The parameterised models add a language part to this. In the Meta-Features model, it is inserted before the rule and object part, making lexical rules language-specific. Due to the modularity of our lexical description framework, language parameters can be inserted before the rule/object part in each module as is illustrated in Figure 14. This makes our fragment more powerful than the abstract model defined in Section 2, which would be equal to allowing a language parameter in the lexeme module but not in any of the other modules. Thus, in our Meta-Features lexicon, we can make rules such as `singular` and `plural`, which are defined in the lexeme space, language-specific, but also phonological rules such as devoicing, which in our fragment only applies to Dutch.

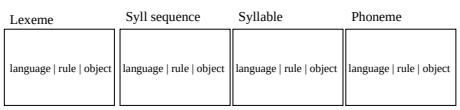


Figure 14: Feature space in the Meta-Features Model

The Meta-Features fragment was implemented using a parallel development strategy. This means that the lexicons for the four languages are developed in parallel and that cross-linguistic generalisations are captured immediately upon construction. This implies that all necessary data is available from the start (which was achieved by implementing the Structure-Sharing fragment first). The same development strategy was used for the Infinitesimal model.

3.2. The Infinitesimal model

We implemented a restricted version of the infinitesimal model. In principle, a language feature can occur anywhere in the feature-value path in the infinitesimal model, at the beginning, at the end, and anywhere in between. In our sample lexicon, a language-feature can be inserted before the lexical rule part and before the object part in each module that is distinguished in our lexical description framework. This situation is illustrated in Figure 15:

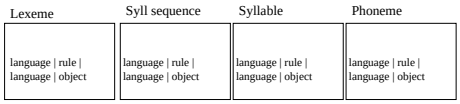


Figure 15: Feature space in the Infinitesimal Model

This way, cross-linguistic generalisations can be captured at the rule and object level in each module, viz. lexeme module, syllable sequence module, syllable module, and phoneme module. For instance, default information in the syllable sequence module can be overridden if the number of syllables that make up a lexical entry in one of the languages is different from the default. For example, the lexeme *Arm* is a monosyllable in Dutch /Arm/, English /A:m/, and Danish /A:m/, but a disyllable in Icelandic /Armur/. This will be represented as follows:

```
Arm:
syll seq = monosyllable
germanic icelandic syll seq = disyllable
```

The default definition of the syllable sequence information is overridden for Icelandic.

3.3. Evaluation

All sample fragments discussed above, cover the same data, but in different ways. Each has its own advantages and disadvantages. The appeal of the Structure-Sharing model is that it provides a rather straightforward way of constructing a multilingual resource from a set of monolingual lexicons using a uniform representation format. A multilingual Structure-Sharing lexicon is constructed by comparing the monolingual hierarchical lexicons for each of the languages and creating a parallel hierarchy containing what the monolingual hierarchies have in common. Apart from being rather straightforward, this procedure can also fairly easily be automated (as described in Cahill (1998)) allowing the automatic construction of lexical resources for NLP applications.

The Structure-Sharing model is also a fairly robust model. Each language has its own hierarchy and language-specific changes can be easily incorporated without affecting the rest of the hierarchy. In the parameterised models, on the other hand, even minor changes can affect the whole hierarchy.

The downside of the Structure-Sharing model is that there is a lot of redundancy. It has more statements per lexical node than the Meta-Features model and the Infinitesimal model[4]. The reason for this is that each language has its own separate hierarchy (or set of hierarchies) and inheritance patterns are repeated over and over again. Even if one would like to add a new dialect (related to one of the languages already available in the lexicon), a complete parallel hierarchy with appropriate links to the parent hierarchy needs to be established. As a consequence, the inheritance network in the Structure-Sharing model might be quite messy and therefore more difficult to maintain and extend.

The parameterised models avoid the kind of redundancy of the Structure-Sharing model. Parameterised multilingual lexicons consist of one single hierarchy in which a language parameter is used to conditionalise certain parts of the hierarchy for certain languages. In our sample lexicons, this language parameter is integrated in DATR's main feature theory which allows us to introduce language variation at different levels in the feature tree – before lexical rules in the Meta-Features model and before lexical rules and object parts in the Infinitesimal model.

Although the Meta-Features model and the Infinitesimal model seem to be able to describe the same data, the Infinitesimal model seems to be preferable as it allows us to capture generalisations that the Meta-Features model could not capture such as cross-linguistic generalisations at the object level in the different modules. However, it is not always self-evident from a linguistic perspective at which levels cross-linguistic generalisations are desirable. By allowing a language parameter to occur before the object part in each of the modules describing the phonological form of a lexeme, semantically related information can be grouped which is not necessarily morphologically and/or phonologically related. In the case of a lexeme such as *Arm*, there is enough morphological and phonological similarity to warrant the approach. Consider, however, the syllable definition for the lexeme *Curve* in English /k3:v/ and Icelandic /hnI:t/. The English /k3:v/ is a CVC syllable with a /k/ onset, a /3:/ peak and a /v/ coda. The Icelandic /hnI:t/, on the other hand, is a CCVC syllable with an onset cluster /hn/, an /I:/ peak and a /t/ as coda. In the Infinitesimal model, all this information can be grouped together in a shared syllable definition. However, do we really want to group this information together? There is no shared element here. More cross-linguistic research could help to define which kinds of cross-linguistic generalisations are linguistically justified.

[4]The repetition of inheritance patterns is even more pronounced with the modular lexical description framework used in the sample fragments. This is because each language has its own lexeme hierarchy, syllable sequence hierarchy, syllable hierarchy, and phoneme hierarchy and for each lexical entry the appropriate values need to be defi ned for all of those for each language in the Structure-Sharing model.

The different ways in which inheritance relations are expressed in the parameterised and non-parameterised models affects the definition of direct interlanguage inheritance. Direct interlanguage inheritance relations are relations where one language inherits characteristics directly from another language such as in the case of borrowings. For example, Dutch borrowed the word *computer* from English and the Dutch lexeme for *Computer* could inherit its phonological form directly from English using interlanguage inheritance relations from English to Dutch. In the non-parameterised Structure-Sharing model, languages inherit shared information from shared hierarchies, and as there is no language feature, there is, in principle, nothing to indicate which hierarchy represents which language. Consequently, modelling direct interlanguage inheritance is not straightforward in this model. In theory, it is possible for the monolingual lexicons to inherit information directly from each other without going via a shared hierarchy, but the actual implementation of a lexicon allowing such inheritance relations is complicated by several engineering issues. Incorporating direct inheritance relations means that the monolingual lexicons are not completely separate anymore. For this to work, one has to make sure that there are no overlapping node names in the different language-specific lexicons, for example, by introducing language identifiers in the node names (as was done in Figure 13). Another side-effect of allowing direct interlanguage inheritance relations is that the resulting multilingual inheritance network becomes messier. There are now several inheritance routes possible for expressing the same shared phenomenon. There is no uniform treatment of interlanguage inheritance anymore. Direct interlanguage inheritance relations can generally be expressed more easily in a parameterised model because all information is integrated into a single hierarchy. In this architecture, a Dutch lexeme could, for example, inherit information directly from English as easily as from Germanic in general.

4. Implications of Results

As noted by Cahill and Gazdar (1999), capturing similarities at different levels of linguistic description in a multilingual inheritance lexicon can contribute significantly to the robustness, maintainability, and extensibility of multilingual NLP systems.

First, a multilingual inheritance architecture offers a more economical encoding of lexical information just as inheritance lexicons in general. As information is stated only once, inheritance lexicons provide the benefit of reduced redundancy and therefore a more concise and transparent storage.

Second, there is the benefit of improved extendability both within languages and to include other, related, languages. It might be possible to add new languages to a lexicon by defining them by difference to related languages already available in the lexicon. For example, Afrikaans could be defined by reference to Dutch similarly to the way Estuary English was defined by reference to English in Section 2.2.2. above.

Third, a multilingual inheritance architecture offers improved robustness. It provides a more intelligent approach to lexical incompleteness. By exploiting default information from both the source and the target language, together with information about the default commonalities across those languages, it may be possible to deduce sufficient information about a missing lexical item via information which is available in the lexicon. Such inferences may not be correct, but they are the best possible guesses that can be made given the way that languages work and given the way they usually relate to each other. For example, the German word for *forbid* could be deduced from the fact that the English verb *bid* translates as *bieten* and that verbs beginning with *for* in English generally begin with *ver* in German. This example is taken from Cahill and Gazdar (1995, p.175).

Finally, a multilingual inheritance lexicon may provide a formal account of how languages have diverged from their common origin. Especially the parameterised models are well-suited for this kind of modelling as they allow us to make a distinction between the different sorts of similarities that exist between languages (i.e. due to genetics, typology, language contact or chance) in the lexical representation. This may be of interest from a linguistic perspective, but it may not be of concern for computational treatments.

5. Conclusion

This paper discussed different architectures for multilingual lexical representation which move away from the traditional Machine Translation architecture to multilingual lexicons. Rather than linking the monolingual lexicons at the level of semantics only, the aim is to encode and exploit lexical similarities between related languages at all levels of linguistic description – morphology, phonology, etc. – by using an inheritance-based formalism.

This paper has shown that at the moment, the question how to build a multilingual inheritance-based lexicon is difficult to answer. It seems that which model is best, depends on what one wants to do with it.

For practical applications, the non-parameterised Structure-Sharing model seems currently the most suitable model. It is relatively straightforward to construct. Each monolingual lexicon keeps its own inheritance structure and shared information is specified in a shared hierarchy from which the monolingual lexicons inherit. There are no preconditions to its construction, i.e. it does not require that all data is available from the start. The disadvantage of the Structure-Sharing model is that there is a lot of redundancy in the model which may make the inheritance network quite messy especially when the network gets bigger.

The construction of a parameterised multilingual lexicon is less straightforward. Parameterised lexicons require more preparatory work. All cross-linguistic data has to be available from the start (which can be quite time-consuming) and in the more powerful models, such as the Infinitesimal model, one has to decide at which levels language variation is allowed in the multilingual lexicon. However, the state of the art in language typology and cross-linguistic research is in general not far enough advanced to guide us in making such decisions. Because of these difficulties, the parameterised models are currently less appealing for practical applications.

From a theoretical perspective, the parameterised models – and in particular the Infinitesimal model – are more interesting than the Structure-Sharing model. As the Infinitesimal model allows us to capture different kinds of generalisations in different ways, it is better placed to provide a linguistic model of the relationships that exist between languages than the other models.

6. Acknowledgements

The research reported here is based on my PhD thesis. I would like to thank my supervisors, Roger Evans, Gerald Gazdar and Adam Kilgarriff, and my examiners, Julie Berndsen and Bill Keller for their valuable comments. Financial support came from the EPSRC and the University of Brighton. Their support is gratefully acknowledged.

7. References

L. Cahill and G. Gazdar. 1995. Multilingual lexicons for related languages. In *Proceedings of the 2nd DTI Language Engineering Conference*, pages 169–176.

L. Cahill and G. Gazdar. 1999. The PolyLex architecture: multilingual lexicons for related languages. *Traitement Automatique des Langues*, 40(2):5–23.

L. Cahill. 1998. Automatic extension of a hierarchical multilingual lexicon. In *Multilinguality in the lexicon II*, pages 16–23, Brighton. ECAI. Workshop at the 13th biennial European Conference on Artificial Intelligence.

A. Copestake, B. Jones, A. Sanfilippo, H. Rodriguez, P. Vossen, S. Montemagni, and E. Marinai. 1992. Multilingual lexical representation. In *ESPRIT BRA-3030 ACQUILEX Working Paper*, number 043. University of Cambridge Computer Laboratory.

W. Daelemans, K. De Smedt, and G. Gazdar. 1992. Inheritance in natural language processing. *Computational Linguistics*, 18(2):205–218.

W. Daelemans and G. Gazdar. 1992. *Special Issues on inheritance*, volume 18.2 & 18.3. Computational Linguistics.

R. Evans and G. Gazdar. 1996. DATR: A language for lexical knowledge representation. *Computational Linguistics*, 22(2):167–216.

R. Evans. 1996. Exploiting inheritance in multilingual lexicons. Paper presented at the AISB-96 Workshop on Multilinguality in the Lexicon, Brighton, also available URL: `http://www.itri.brighton.ac.uk/~Roger.Evans/papers/aisb96`.

U. Heid and K. Krüger. 1996. A multilingual lexicon based on frame semantics. In *Proceedings of the AISB-96 Workshop on Multilinguality in the Lexicon*, pages 1–13, Brighton.

M. Kameyama. 1988. Atomization in grammar sharing. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

A. Kilgarriff, L. Cahill, and R. Evans. 1999. The GREG framework for multilingual valency lexicons. GREG deliverable 2.1, ITRI.

C. Tiberius and R. Evans. 2000. Phonological feature based multilingual lexical description. In *Proceedings of TALN 2000*, pages 347–356, Lausanne, October.