The Treatment of Derivational Morphology in a Multilingual Transfer-based MT System (Eurotra)

S. Ananiadou, A. Ralli and A. Villalva

In this paper, we examine the impact of using interlingual labels in derivational morphology. We demonstrate that interlingual labels in derivational morphology contribute to reducing complexity in transfer. Our work is based on languages that belong to different language families, i.e. English, Greek and Portuguese, and was carried out in the context of the Eurotra project.

1. The Eurotra Model

The Eurotra model (ET) is a transfer based Machine Translation system consisting of several representational levels, both in analysis and synthesis, i.e. a morphological level (EMS), a surface phrasal syntax level indicating order of constituents (ECS), a relational syntax level with frame information (ERS), and an interface structure (IS), which is a deep syntactic dependency representation, consisting of governor–argument–modifier structures and featurised morphosyntactic and semantic information.

Each level consists of a dictionary, a generator and a translator. The dictionary captures the relevant information, expressed by means of features (attribute/value pairs), for each lexical unit. The generator reflects information concerning grammatical structures and translators link contiguous levels.

From a translational point of view, IS is the most crucial level in ET, as this is the level at which mappings between the source and the target languages take place. EMS, ECS and ERS reflect surface and relational

morphosyntactic language-dependent relations. The IS level neutralises language-dependent information, in several ways, e.g., by using a canonical geometry.

One of the main notions underlying the ET philosophy is simple transfer. This notion relates to the fact that ET deals with a multilingual environment (nine languages, seventy two transfer modules) where complex transfer should be avoided, where possible. By simple transfer we mean correspondence of lexical equivalents. Complex transfer triggers the mapping of a structure into a different structure, thus, the necessity of writing transfer structural rules.

In order to reduce complex transfer, a set of linguistic phenomena, such as tense and aspect, diathesis, determination and derivational morphology, among others, is treated using interlingual labels, at IS, even though ET is a transfer based MT system. Our concern in this paper is with the interlingual treatment of derivational morphology.

2. Definition of Derivation

Derivation is, along with compounding and inflection, a word formation process, i.e., it consists of a set of linguistic operations applied either to a stem, to a root or to an already existing word, to form another word. These operations also allow a derived word to be interpreted by analysing its internal structure.

We define a derivational operation in terms of the association of one categorial relationship with one semantic operation and a morphological one. In order to perform this derivational operation it is necessary to have access to:

(1) A. categorial relationship

1. the grammatical category of the base

   In Portuguese, the suffix -'mento' selects verbs:

   arrefecer → arrefecimento

   'to get cold' → 'the process of getting cold'

2. the grammatical category of the affix

   In Greek the noun forming suffix -'$\sigma\eta$' belongs to the category of

nouns:

επιταχυνω → *επιταχυνση*[1]

'to speed up' → 'the process of speeding up'

3. the morphosyntactic subcategories of the base(e.g. gender, argument structure, etc.)

   In Portuguese the suffix -'mente' selects the feminine form of the adjective:

   espantosa → espantosamente

   'astonishing' → 'astonishingly'

4. the morphosyntactic subcategories of the affix(e.g. gender, inflectional class, etc.)

   In Greek the agentive affix bears information relative to gender:

   -'*τηs*' for masculine: *παικτηs* 'player'

   -'*τρια*' for feminine: *παικτρια* 'female player'

5. the subcategorisation features of the affix i.e. information concerning the selection of a specific base with respect to its grammatical category.

   The Greek suffix -'*μα*' selects verb bases:

   περνω→*περασμα* 'pass'→'passage'.

B. semantic operation

1. semantic features of the base, such as [± animate]
2. semantic function of the affix, such as [± animate]

C. morphological operation

1. type of affix (e.g. level ordering affixation, cf. Selkirk (1982))
2. diacritics (e.g.[± latinate] for English words borrowed from Latin or French) (cf. Aronoff (1976))

Each affix is related to only one derivational operation. However, it is possible that the same affix is added to different bases (cf.(2)) and also the same derivational process can be obtained by means of combining different affixal forms to categorically distinct bases (cf.(3)).

In Portuguese, the same suffix creating action nouns is attached to different category bases. It subcategorizes either verb stems or adjectives:

[1] In most Greek examples derivational affixes are followed by inflectional endings; in '*επιταχυνση*' the derivational suffix '*ση*' includes the inflectional ending '*η*'.

(2) formaliza → formalizão 'formalisation'
erudito → erudião 'erudition'
formalise → formalisation
transport → transporter

In Greek, the agentive noun formation is realized by two different suffixes : -'τηs' and -'εαs'. The first suffix selects verbs while the second attaches to noun stems:

(3) χτιζω → χτιστηs 'builder'
μεταφορα → μεταφορεαs 'transporter'

This type of phenomena is usually the result of language change. The superposition of several layers of derivational operations resulting in the production of derived words, apparently similar in structure and meaning, makes it difficult to determine the appropriate template for each affix, which is crucial for the use of abstract labels. This is a task that has to be carefully fulfilled by individual language grammer writers.

3. Treatment of Derivational Morphology in Eurotra

In this section we will examine the linguistic information attached to a string/lexical unit, the way derivations are represented and how they are built, in each representational level.

After initial segmentation (ET front-end), strings are input to the subsequent morphological level (EMS). The dictionary of EMS consists of affixes, roots, stems and words. The minimal set of features needed is as follows:

(4) mc (morphological class)=suffix, prefix, root, stem, word.
bar=zero, minus.
cat=n(noun), adj(adjective), v(verb), adv(adverb).
level=one, two.
subcat(subcategorization)=n, adj, v.
string=.
lu=.

The 'mc' attribute identifies the morphological class of a given dictionary entry. The 'bar' attribute is used according to existing linguistic theories

(Selkirk (1982)) which stipulate that the word is an X^0 category (a lexical category). The 'bar=minus' feature ensures that roots, stems and affixes do not occur independently without attaching to a base. The attribute 'level' characterizes affixes in terms of two different levels depending on selective restriction; + ity(level one) attaches to roots 'activ-ity', while #ness(level two) attaches to words 'tender-ness'. The feature concerning subcategorization 'subcat' stipulates the type of category an affix is attached to e.g. the suffix +ity attaches to adjectives. The value of 'string' corresponds to the surface realisation of an affix, root, stem or word i.e., string={ity, activ}. The 'lu' value corresponds to the basic form.

The abstract label for action nouns has the following dictionary entry:

(5) l_act_n={cat=n, bar=minus, lu=#act_n, string=(age;ation; ment;al;ing)}

Each affix has a separate entry specifying additional information:

(6) l_ity={mc=suffix, bar=minus, cat=n, level=one, subcat=adj, string=ity, lu=ity}.

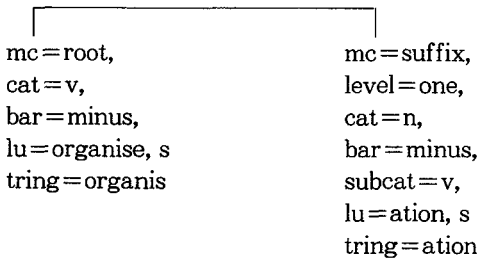
The word 'activity' is built by the following structure building rule:

(7) b_ity={cat=n, bar=zero, mc=word} [{cat=adj, mc=root}, {cat=n, mc=suffix, string=ity, lu=ity}].

We know from the dictionary that the suffix -'ity' has a set of properties which allow it to attach only to roots, thus the string *happyity will not be accepted. Similarly, we can block other unwanted analyses e.g. *activness will not be permitted, as #ness attaches only to words.

Word structure is maintained at all levels, except when lexicalisation occurs (cf. section 4). At EMS, we have the following represention for the word 'organisation':

(8) mc=word, bar=zero, cat=n



At ERS and IS, the word structure is a dependency structure. The head of a derived word (typically the suffix) is the governor, first daughter of the structure; the base is an argument and, in general, prefixes which are not category changing, are modifiers.

The corresponding representation at IS is similar for the three languages. This similarity makes obvious the use of incorporating morphological analysis and, more specifically, the advantages of using abstract labels in transfer.

(9) mc=word, bar=zero, cat=n

role=gov,	role=arg 1,
cat=n,	cat=v
lu=#act_n	lu=(organise;organiza; orghanwnw)

Frame inheritance of affixes is currently under investigation in the project. For example the suffix -'er' as in drink→drinker blocks the agent but allows the theme. Frame information could be added in the dictionary.

4. Compositionality and Lexicalisation

In morphology, compositionality is related to the possibility of identifying the internal structure of complex words, i.e., the word can be decomposed in its constituent elements which are morphological entities (roots, stems and affixes).

Even though the internal structure of a word may be successfully identified, it is compositional only if the semantic interpretation can be calculated from its constituent elements. This is the case with the Greek word *κινηση* 'movement', derived form the verb *κινω* 'to move'.

Moreover, structural compositionality is related to the presence or absence of systematic morphographemic changes (excluding regular allomorphy graphically expressed and regular truncation operations) occurring during the derivational processes. If unpredictable morphographemic changes occur, derived words are non-compositional. In Portuguese, the word construção is derived from the verb construir 'to construct'. The compositional form would be *construição. If we compare this derived word

with others sharing the same root, we conclude that the absence of the thematic vowel -i- is absolutely unpredictable (cf. destruir → destrução).

Lexicalisation occurs when a derived word cannot be decomposed, therefore it is not compositional. Lexicalisation is a destructive operation in the sense that substructure of the word is deleted. The relevant information is percolated to the top node and the entry is representational level. Specific translator rules (t-rules) take care of lexicalising a word from level to level where appropriate.

5. Abstract Labels

Following work done by Sproat (1985) and Corbin (1988), Verheul (1989) suggested the use of a set of abstract labels in ET: "These labels are intended to be language independent, that is, all Eurotra languages use the same set of interlingual labels. This entails that the translation of affixes can be done by default rule. During analysis and generation the translation from affix into abstract label and vice versa will have to take place. The n-to-one relation between an affix and a meaning operation has to be handled in the monolingual component"(Eurotra Working Papers).

Translational evidence shows that there is no one-to-one relation between affixes among languages.

(10) 1. English → Greek
writer → συγγραφεας
driver → οδηγοss
employer → εργοδοτης
2. English → Portuguese
listener → ouvinte
player → jogador
3. Greek → Portuguese (or to English)
αξιολογηση → avaliação → evaluation
4. Portuguese → English
estudante → student
ouvinte → listener
5. Portuguese → Greek
apresentão → παρουσιαση 'presentation'
adequão → καταλληλοτητα 'adequacy'

The designation of each abstract label is adopted according to its general meaning and to the category of the derived word. The relations between abstract labels and the way they are expressed in all Eurotra languages should take into account the types of derivation involved and the argument structure of the derived word.

Verheul's proposal included a small set of abstract labels for deverbals. We will now present a larger set (although not a complete one), which also includes abstract labels for other types of derivation. The first subset concerns nominalisation processes. The abstract labels that we propose are:

(11) 1. #ag_n for agential nouns. This label correponds to suffixes that attach to verb bases:

En. observe → observer

Gr. παρατηρω → παρατηρητης 'observe → observer'

Pt. observado → observador 'observed → observer'

2. #act_n for action, process or result nouns. This label corresponds to suffixes that attach to verb bases:

En. invest → investment

Gr. επενδυω → επενδυση 'invest → investment'

Pt. investi → investimento 'invest → investment'

3. #state_n for nouns that express a property. This label corresponds to suffixes that attach to adjective bases:

En. equal → equality

Gr. ισος → ισοτητα 'equal → equality'

Pt. igual → igualdade 'equal → equality'

4. #dim_n for diminutive nouns. This label corresponds to suffixes that attach to noun bases:

En. dog → doggie

Gr. σπιτι → σπιτακι 'house → little house'

Pt. mesa → mesinha 'table → small table'

5. #aug_n for augmentative nouns. This label corresponds to suffixes that attach to noun bases:

Gr. πορτα → πορταρα 'door → big door'

Pt. porta → portão 'door → big door'

6. #repet_n for nouns expressing repetion. This label corresponds to prefixes that attach to deverbal nouns:

En. analysis → reanalysis
Pt. constrûo → reconstrûo
Gr. *επαναπροσδιορισμος* → redefinition 'construction → reconstruction'

The second subset includes adjectivalisation procosses. The abstract labels that we propose are the following:

(12) 1. #poss_adj for adjectives expressing a possibilty. This label corresponds to suffixes that attach to transitive verbs:
En. translate → translatable
Gr. *μεταφραση* → *μεταφρασιμος* 'translate → translatable'
Pt. traduzir → traduzível 'translate → translatable'
2. #rel_adj for relational adjectives. This label corresponds to suffixes that attach to nouns:
En. structure → structural
Gr. *κυβερνηση* → *κυβερνητικος* 'government → governmental'
Pt. forma → formal 'form → formal'
3. #intens_n for adjectives expressing intensification. This label corresponds to affixes attach to adjectives:
En. active → hyperactive
Gr. *καντικος* → *υπερκαντικος* 'active → hyperactive'
Pt. care → caríssimo 'expensive → very expensive'
4. #neg_adj for adjectives expressing negation. This label corresponds to prefixes that attach to adjectives:
En. happy → unhappy
Gr. *γνωστος* → *αγνωστος* 'known → unkown'
Pt. útil → inútil 'useful → unuseful'
5. #repet_adj for adjectives expressing repetition. This label corresponds to prefixes that attach to deverbal adjectives:
En. usable → reusable
Gr. *γραμμενος* → *ξαναγραμμενος* 'written → rewritten'
Pt. utilizável → reutilizável 'usable → reusable'

The third subset includes verbalisation processes. The abstract labels that we propose are the following:

(13) 1. #make_v for causative verbs. This label corresponds to suffixes

that attach either to nouns or to adjectives:

En. drama → dramatise

Gr. δραμα → δραματοποιω 'drama → dramatise'

Gr. μεγαλος → μεγαλωνω 'big → make big'

Pt. moral → moralizar 'moral → moralise'

2. #repet_v for verbs expressing a repetition. This label correponds to prefixes that attach to verbs:

   En. translate → retranslate

   Gr. γραφω → ξαναγραφω 'write → rewrite'

   Pt. fazer → refazer 'do → redo'

3. #oppos_v for reversative verbs. This label correponds to prefixes that attach to verbs:

   En. connect → disconnect

   Gr. συνδεω → αποσυνδεω 'connect → disconnect'

   Pt. ligar → desligar 'connect → disconnect'

The examples that we presented show that, in some cases, one of the languages makes use of another linguistic strategy, namely a syntactic strategy, to express a given meaning, which, in other languages, is morphologically encoded. This is the case of #aug_n. Therefore for certain mappings of abstract labels we do need to use complex transfer, however in general simple transfer is possible for most mappings involving an abstact label.

The use of abstract labels in monolingual analysis is quite straightforward, but, in systhesis, it implies a thorough knowledge of the morphological properties of each particular affix and each specific derivational process.

6. Conclusion

This paper presents ongoing research in morphology in Eurotra. Our main purpose is to facilitate transfer using elegant linguistic solutions as well as computationally efficient ones. Abstract labels provide an interesting mechanism in treating compositionally derivational morphology as the examples given from three different languages have demonstrated.

7. Acknowledgement and Disclaimer

The research reported on here was undertaken in the framework of the Eurotra Machine Translation Project, sponsored mainly by Commission of the European Communities. We thank J. McNaught, UMIST, for proof reading this paper. The views of the authors of this paper are not necessarily those of the Eurotra Project Management.

References

Allen, M. R. (1978) 'Morphological Investigation,' Ph. D. Dissertation, University of Connecticut.

Ananiadou, S., Ralli, A., Villalva, A. (1990) 'Derivational Morphology,' *Eurotra Final Research Report on Word Structure.*

Ananiadou, S. & Verheul, C. (eds.) (forthcoming) 'Morphology in Eurotra,' *Eurotra Papers in Machine Translation and Natural Language Processing*, Volume 3, DG XIII, CEC, Luxembourg.

Aronoff, M. (1976) *Word Formation in Generative Grammar*, MIT Press, Cambidge, Mass.

Bech, A. & Nygaard, A. (1988) 'The E-framework: A Formalism for Natural Language Processing,' In Proceedings of the 12th International Conference on Computational Linguisitcs, *COLING '88*, Budapest.

Corbin, D. (1987) *Morphologie Dérivationnelle et Structuration du Lexique 2 vol*, Tübingen: Max Niemeyer Verlag.

Corbin, D. (1989) 'Form, Structure and Meaning of Constructed Words in an Associative and Stratified Lexical Component,' In *Yearbook of Morphology* 2, 31-54.

Di Sciullo, A. M. & Williams, E. (1987) *On the Definition of Word*, MIT Press, Cambridge, Mass.

Everaert, M., Everts, A., Huybregts, R. & Tromelen, M. (eds.) (1988) *Morphology and Modularity*, Foris, Dordrecht.

Selkirk, E. (1982) *The Syntax of Words*, MIT Press, Cambridge, Mass.

Sproat, R. (1985) 'On Deriving the Lexicon,' Ph. D. Dissertation, MIT.

Verheul, C. (forthcoming) 'Derivation,' in Ananiadou, S. & Verheul, C. (eds.).

Dr. S. Ananiadou
Centre for Computational Linguistics
UMIST, Manchester
England

Dr. A. Ralli
Department of French
University of Athens
Greece

Dr. A. Villalva
Department of Linguistics
University of Lisbon
Portugal