Multi-Syllable Phonotactic Modelling

Anja Belz
CCSRC, SRI International
23 Millers Yard, Mill Lane
Cambridge CB2 1RQ, UK
`anjab@cam.sri.com`

Abstract

This paper describes a novel approach to constructing phonotactic models. The underlying theoretical approach to phonological description is the multi-syllable approach in which multiple syllable classes are defined that reflect phonotactically idiosyncratic syllable subcategories. A new finite-state formalism, OFS Modelling, is used as a tool for encoding, automatically constructing and generalising phonotactic descriptions. Language-independent prototype models are constructed which are instantiated on the basis of data sets of phonological strings, and generalised with a clustering algorithm. The resulting approach enables the automatic construction of phonotactic models that encode arbitrarily close approximations of a language's set of attested phonological forms. The approach is applied to the construction of multi-syllable word-level phonotactic models for German, English and Dutch.

1 Introduction

Finite-state models of phonotactics have been used in automatic language identification (Zissman, 1995; Belz, 2000), in speech recognition (Carson-Berndsen, 1992; Jusek et al., 1994; Jusek et al., 1996; Carson-Berndsen, 2000), and optical character recognition, among other applications. While statistical models (n-gram or Markov models) are derived automatically from data, their symbolic equivalents are usually constructed in a painstaking manual process, and — because based on standard single-syllable phonological analyses — tend to overgeneralise greatly over a language's set of wellformed phonological strings. This paper describes methods that enable the automatic construction of symbolic phonotactic models that are more accurate representations of phonological grammars.

The underlying theoretical approach to phonological description is the *Multi-Syllable Approach* (Belz, 1998; Belz, 2000). Syllable phonotactics vary considerably not only in correlation with a syllable's position within a word, but also with other factors such as position relative to word stress. Analyses based on multiple syllable classes defined to reflect such factors can more accurately account for the phonologies of natural languages than analyses based on a single syllable class.

Object-Based Finite State Modelling (previously described in Belz, 2000) is used as an encoding, construction and generalisation tool, and facilitates *Language-Independent Prototyping*, where incompletely specified generic models are constructed for groups of languages and subsequently instantiated and generalised automatically to fully specified, language-specific models using data sets of phoneme strings from individual languages. The theory-driven (manual) component in this construction method is restricted to specifying the maximum possible ways in which syllable phonotactics may differ in a family of languages, without hard-wiring the differences into the final models. The actual construction of models for individual languages is a data-driven process and is done automatically.

Sets of German, English and Dutch syllables were used extensively in the research described in this paper, both as a source of evidence in support of the multi-syllable approach (Section 2) and as data in automatic phonotactic model construction (Section 4). All syllable sets were derived from sets of fully syllabified, phonetically transcribed forms collected from the lexical database CELEX (Baayen et al., 1995). CELEX contains compounds and phrases as well as single words. Phonological words were defined as any phonetic sequence with a single primary stress marker, and all other entries were disregarded.

2 Multi-Syllable Phonotactics

The multi-syllable approach works on the assumption that single-syllable approaches cannot adequately capture the phonological grammars of natural languages, because they fail to account for the significant syllable-based phonotactic variation resulting from a range of factors that is evident in natural languages, and consequently overgeneralise greatly.

Single-syllable analyses. The traditional view is that all syllables in a language share the same structure and compositional constraints which can

	German		English		Dutch	
	all	unique (%)	all	unique (%)	all	unique (%)
Initial	3,806	624 (16.4%)	6,177	2,657 (43.01%)	5,476	947 (17.29%)
Medial	3,832	358 (9.34%)	3,149	344 (10.92%)	5,446	723 (13.28%)
Final	7,040	2,133 (30.3%)	6,750	2,132 (31.59%)	7,279	1,786 (24.54%)
Monosyllables	5,114	855 (16.72%)	7,265	2,963 (40.78%)	5,641	718 (12.73%)
TOTAL	10,606	3,970 (37.43%)	14,333	8,096 (56.49%)	11,448	4,174 (36.46%)

Table 1: Syllable set sizes and number of syllables unique to each set (position).

be captured by a single analysis. In many languages, however, the sets of word-initial and/or word-final consonant clusters differ significantly from other consonantal clusters (Goldsmith, 1990, p. 107ff, lists several examples from different languages). Such idiosyncratic clusters have been treated as 'terminations', 'appendices', or as 'extrasyllabic' (Goldsmith, 1990), and integrated along with syllables at the word-level. Similar, apparently irregular phenomena occur in correlation with tone and stress, and the first and last vocalic segments in phonological words are often analysed as 'extratonal' and 'extrametrical'. However, such apparent irregularities are not restricted to the beginnings and ends of phonological words, and the phonotactics of syllables are affected by a range of factors other than position, which are difficult if not impossible to account for by the notion of extrasyllabicity.

Three problematic issues arise in single-syllable analyses. Firstly, if a phonotactic model assumes a single syllable class for a language, and if the language has idiosyncratic word-initial and word-final phonotactics, then the set of possible phonological words that the model encodes is necessarily too large, and includes words that form systematic (rather than accidental) gaps in the languages. Secondly, if extrasyllabicity is used to account for phonotactic idiosyncracies, then the resulting theory of syllable structure fails to account for everything that it is intended to account for, and is forced to integrate constituents that are not syllables (the *extra*syllabic material) at the word level. Thirdly, the notion of extrasyllabicity only works for cases where phonemic material can be segmented off adjacent syllables (most easily done at the beginnings and ends of words), and cannot be used to account for syllable-internal variation. The alternative offered by multi-syllable analyses is to make the universal assumption that position, stress and tone (among other factors) will result in variation in syllable phonotactics that are not necessarily restricted to any particular part of words, and to account for such variation systematically by the use of different syllable classes.

Related approaches. The idea to discriminate between different syllable types, classified by word position and position with respect to the stressed syllable has been explored and utilised in previous research, for example in FSA-based phonotactic models, typed formalisms, and in stochastic production rule grammars. Carson-Berndsen (1992) uses two separate FSAs to encode the phonotactics of full and reduced syllables, and Jusek et al. (1994) distinguish between stressed and unstressed syllables. In a typed feature system of morpho-phonology, Mastroianni and Carpenter (1994) define subtypes of the general type *syllable*.

The most closely related existing research is that presented by Coleman and Pierrehumbert (1997). The paper examines different possibilities for using a probabilistic grammar for English words to model native speakers' acceptability judgments. The production rule grammar encodes the phonotactics of English monosyllabic and bisyllabic words. Different probability distributions over paths in derivation trees are investigated which model likelihood of acceptability to native speakers, rather than likelihood of occurrence. To build a grammar that accounts for interactions among onsets and rhymes, location with respect to the word edge and word stress patterns, six syllable types are distinguished which reflect possible combinations of the features strong, weak, initial and final. The subsyllabic constituents onset and rhyme are similarly marked for stress and position.

The present research extends existing work on syllable subclasses by applying the multi-syllable approach systematically to model the entire phonotactics of languages, and by using it for language-independent prototyping (see Section 3.3 below).

Position-correlated phonotactic variation. Table 1 shows statistics for sets of monosyllabic words and initial, medial and final syllables in CELEX. For each language and each syllable set, the table shows the size of the set (e.g. there are $3,806$ different initial German syllables in CELEX), and the size of its subset of syllables that do not occur in any other set (e.g. 624 out of $3,806$ initial German syllables, or 16.4%, only occur word-initially). For all three languages, the figures show significant differences between the sets of syllables that can occur in the four different positions and their unique

		Medial	Final	Mono
German:	Initial	2,619 (0.52)	1,466 (0.16)	1,392 (0.18)
	Medial		1,928 (0.22)	1,185 (0.15)
	Final			3,873 (0.47)
English:	Initial	1,860 (0.25)	1,920 (0.17)	2,266 (0.20)
	Medial		1,787 (0.22)	1,008 (0.11)
	Final			3,576 (0.34)
Dutch:	Initial	3,594 (0.49)	2,764 (0.28)	3,003 (0.37)
	Medial		3,279 (0.35)	2,428 (0.28)
	Final			4,320 (0.50)

Table 2: Intersections and set similarities for German, English and Dutch syllables (position).

	German		English		Dutch	
	all	unique (%)	all	unique (%)	all	unique (%)
Stressed	8,919	2,977 (33.37%)	9,399	5,280 (56.18%)	9,934	3,484 (35.07%)
Pretonic	989	30 (3.03%)	3,201	1,362 (42.55%)	1,780	71 (3.99%)
Posttonic	5,897	388 (6.58%)	4,754	670 (14.09%)	5,960	517 (8.67%)
Plain	6,819	229 (3.36%)	6,020	944 (15.68%)	6,662	176 (2.64%)
TOTAL	10,598	3,624 (34.20%)	14,333	8,256 (57.60%)	11,443	4,248 (37.12%)

Table 3: Syllable set sizes and number of syllables unique to each set (stress).

subsets. In German and Dutch, final syllables are particularly idiosyncratic, with 30.3% and 24.54%, respectively, not occurring in any other position. In English, all syllable sets except the medial syllables display a high degree of idiosyncracy. Table 2 shows the size of the intersections between the syllable sets, and the more objective measure of set similarity in brackets[1]. In German and Dutch, the similarity between initial and medial syllables, and between final and monosyllables is particularly high. The similarity between the least similar of syllable sets is much greater in Dutch than in either English or German. In English, only the final and monosyllables display any significant similarity. Average set similarity is highest in Dutch (0.37), followed by German (0.28), and English (0.21).

Stress-correlated phonotactic variation. Table 3 shows analogous statistics for phonotactic variation correlated with word stress. Set sizes and unique subset sizes are shown for the set of syllables that carry primary stress (stressed), those immediately preceding stress (pretonic), those immediately following stress (posttonic), and all others (plain). In all three languages, the set of stressed syllables has least in common with other sets. In English, this is closely followed by the pretonic syllables. The average percentage of syllables unique to a set is highest in English, followed by Dutch and then German.

These statistics show not only that there is significant syllable-level variation in the phonotactics of all three languages, but also that the simple strategy of subdividing the set of all syllables on the basis of position and stress succeeds in capturing at least some of this variation. If a high percentage of syllables in one subcategory do not occur in any other, then distinguishing this syllable subcategory in a phonotactic model will help reduce overgeneralisation.

3 Encoding, Construction and Generalisation of Phonotactic Models

3.1 Object-Based Finite-State Modelling

The OFS Modelling formalism was used as a tool for encoding, constructing and generalising phonotactic models in the research described in Section 4. OFS Modelling consists of three main components, (i) a representation formalism, (ii) a mechanism for automatic model construction, and (iii) mechanisms for model generalisation. Brief summaries of the components that were used in the research described in this paper are given here (for full details see Belz, 2000).

Underlying OFS Modelling is a set of assumptions about linguistic description that shares many of the fundamental tenets of declarative phonology (Bird, 1991, for example). This set of assumptions includes a strictly non-derivational, non-transformational and constraint-based approach to linguistic description, and the principle of constraint inviolability.

---

[1] Set similarity here is the standard measure of the size of the intersection over the size of the union of two sets S_1 and S_2, or $|S_1 \cap S_2|/|S_1 \cup S_2|$ (not defined for $S_1 = S_2 = \emptyset$).

The OFS formalism is a declarative, monostratal finite-state representation formalism that is intuitively readable, facilitates the automatic data-driven construction of models, and permits the integration of available prior, theoretical knowledge. The derivations (trees or brackettings) defined by OFS models correspond to context-free derivations with a limited tree depth or degree of nesting of brackets. This means that in OFS models (unlike in other normal forms for regular grammars), rules (hence expansions or brackets) can, if appropriately defined, systematically correspond to standard linguistic objects, the reason why the formalism is called *object-based*.

OFS Model $O = (N, T, P, n+1)$			
n:	O_0^n	$\Rightarrow$	ω_0^n
n-1:	O_0^{n-1}	$\Rightarrow$	ω_0^{n-1}
	O_1^{n-1}	$\Rightarrow$	ω_1^{n-1}
	...		
	O_m^{n-1}	$\Rightarrow$	ω_m^{n-1}
...			
1:	O_0^1	$\Rightarrow$	ω_0^1
	O_1^1	$\Rightarrow$	ω_1^1
	...		
	O_l^1	$\Rightarrow$	ω_l^1
0:	O_0^0	$\Rightarrow$	ω_0^0
	O_1^0	$\Rightarrow$	ω_1^0
	...		
	O_p^0	$\Rightarrow$	ω_p^0

Figure 1: Notational convention for OFS models.

OFS Models. The OFS representation formalism is essentially a normal form for regular sets. OFS models can be interpreted in the same way as standard production rule grammars, but are subject to a set of additional constraints. An OFS model O is denoted $(N, T, P, n+1)$, where N is a finite set of non-terminal objects O_j^i, $0 \leq i \leq n$, and T is a finite set of terminals. P is an ordered finite set of n sets of productions $O_j^i \Rightarrow \omega_j^i$, where $O_j^i \in N$, and for $i > 0$, ω_j^i is a regular expression[2] over symbols $O_h^g \in N, i > g$, whereas for $i = 0$, ω_j^i is a set of strings[3] from T^*. An OFS model O has n levels, or sets of production rules, and each rule $O_j^i \Rightarrow \omega_j^i$ is uniquely associated with one of the levels. The nth set of production rules is a singleton set $\{O_0^n \Rightarrow \omega_0^n\}$, and O_0^n is interpreted as the start symbol. The notational convention adopted for OFS models is as shown in Figure 1.

Definition 1 *OFS Model*
An OFS model O is a 4-tuple $(N, T, P, n+1)$, where N is a finite set of nonterminals O_j^i, $0 \leq i \leq n$, $O_0^n \in N$ is the start symbol, T is a finite set of terminals, $n+1$ denotes the number of levels in the model, and $P =$

$$\{\ \{O_0^n \Rightarrow \omega_0^n\},$$
$$\{O_0^{n-1} \Rightarrow \omega_0^{n-1},\ O_1^{n-1} \Rightarrow \omega_1^{n-1},\ \ldots O_m^{n-1} \Rightarrow \omega_m^{n-1}\},$$
$$\ldots$$
$$\{O_0^1 \Rightarrow \omega_0^1,\ O_1^1 \Rightarrow \omega_1^1,\ \ldots O_l^1 \Rightarrow \omega_l^1\},$$
$$\{O_0^0 \Rightarrow \omega_0^0,\ O_1^0 \Rightarrow \omega_1^0,\ \ldots O_p^0 \Rightarrow \omega_p^0\ \}\ \},$$

where each rule $O_j^i \Rightarrow \omega_j^i$ is uniquely associated with one of the levels, ω_j^0 is a set of strings from T^*, $\omega_j^i, i > 0$, is a regular expression over objects $O_h^g \in N, i > g$.

Each rule $O \Rightarrow \omega$ in an OFS model corresponds to a set of strings which will be referred to as an object set or class, where O is the name of the object. The production rules in OFS models will also be referred to as object rules.

OFS models thus differ from standard production rule grammars in three ways. Firstly, RHSs of rules above level 0 are arbitrary regular expressions[4]. Secondly, terminals from T are restricted to appearing in the RHSs of rules at level 0 (mostly to facilitate automatic model construction, see below). Thirdly, OFS models are limited in their representational power to the finite-state domain by the constraints that the RHSs of rules in rule sets at level $i > 0$ are regular expressions over non-terminals that appear only in the LHSs of rules in rule sets at levels $g < i$. That this limits representational power to the regular languages can be seen from the fact that all non-terminals O_j^i in the RHS of the single top-level rule can be substituted iteratively with the RHSs of the corresponding rules $O_j^i \Rightarrow \omega_j^i$. This iteration terminates after a finite time because there is a finite number of levels in the model, and at this point the RHS of the top-level rule contains only non-terminals, i.e. *is* a regular expression, hence represents a regular language.

Unlike other normal forms for regular production-rule grammars (such as left-linear and right-linear

---

[2] In the regular expressions in this paper, r^* denotes any number of repetitions of r, r^+ denotes at least one repetition of r, and $r + e$ denotes the disjunction of r and e.

[3] The string sets in level 0 RHSs are actually implemented more efficiently as finite automata.

[4] Other formalisms for linguistic analysis have permitted full regular expressions in the RHSs of rules. For instance, in syntactic grammars, the recursive nature of some types of coordination has been modelled with right-recursive regular expressions (e.g. in GPSG).

sets of production rules), OFS models enable the definition of production rules and hence derivations that can, if appropriately defined, correspond to standard linguistic objects and constituents (not possible in linear grammars). Through the association of rules with a finite number of levels, OFS models permit the definition of grammars that encode sets of context-free derivations up to a maximum depth equal to the number of levels in the model.

The fact that non-terminal strings are in OFS models restricted to the lowest level, facilitates the combined theory and data driven construction of models. Uninstantiated models can be defined, that encode what is known in advance about the structural regularities of the object to be modelled in levels above 0, and have under-specified level 0 RHSs that are subsequently instantiated on the basis of data sets of examples of the object to be modelled. OFS Modelling also has a generalisation procedure which can be used to generalise fully instantiated OFS models. Each of these mechanisms is described in turn over the following paragraphs.

Uninstantiated OFS Models. In fully specified OFS models (as defined in the preceding section), the right-hand sides (RHSs) of production rules at level i are regular expressions for $i > 0$, and string sets for $i = 0$. This separation makes it simple to construct incompletely specified models, or *prototype OFS models*, where the RHSs of level 0 rules are pattern descriptions rather than strings sets. Level 0 RHSs in prototype models have the form $O_i^0 \Rightarrow S_i$, where O_i^0 is the name of the object, and S_i is a set former $\{x : \mathbf{v}x\mathbf{w} \in D, P_1, P_2, \ldots P_n\}$, where $\mathbf{v}, \mathbf{w}$ are concatenations of variables, D refers to any given finite data set of strings, and $P_i, 1 \leq i \leq n$ are properties of the variables in $\mathbf{v}$ and $\mathbf{w}$.

Instantiation of Prototype OFS Models. The OFS instantiation procedure takes a prototype OFS model M for some linguistic object and a data set D of example members of the corresponding object class and proceeds as follows. For each level 0 rule $O_i^0 \Rightarrow S_i$ in M, and for each element x of D, all substrings of x that match S_i are collected. The resulting set of substrings becomes the new RHS of rule O_i^0. After instantiation, level 0 rules whose RHS is the empty set are removed, as are rules at higher levels whose RHSs contain non-terminals that can no longer be expanded by any of the production rules in M.

Object-Set Generalisation. Instantiated OFS models can be generalised by object-set (OS) generalisation, where pairs of level 0 object sets are compared on the basis of a standard set similarity measure *sim* for two finite sets D_1 and D_2 (not defined for $D_1 = D_2 = \emptyset$): $sim(D_1, D_2) = |D_1 \cap D_2|/|D_1 \cup D_2|$. The OS-generalisation procedure takes a fully specified OFS model M and a given similarity threshold τ, and, applying a simple clustering algorithm, merges all object sets that have a similarity value *sim* matching or exceeding τ. That is, the OS-generalisation procedure measures the similarity between all pairs of level 0 sets, and all pairs that match or exceed the threshold end up in the same cluster. Finally, the old object names (non-terminals) in the RHSs of object rules at levels above 0 are replaced with the LHSs of the corresponding new merged object rule, while all object rules that now have identical RHSs are in turn merged. In this way, generalisation 'percolates' upwards through the levels of the model.

Determining an appropriate value for the similarity threshold τ is not unproblematic. It could be set in relation to the average similarity value in an instantiated model (individually for each prototype instantiation), but this approach would obscure the similarities that object-set generalisation (in particular in conjunction with LIP) is intended to exploit. The whole point of object-set generalisation for language-independent prototypes is that it will merge a different number of level 0 object classes in different prototype instantiations, creating different final, language-specific OFS models. If τ is set in proportion to the average similarity between level 0 classes, then this difference is reduced, and the resulting models will tend to retain the same number of level 0 object classes from the prototype. For example, if the above prototype model $Word$ is instantiated to a data set from a language that has phonotactics which differ only between stressed and unstressed syllables, then all similarity values between stressed syllable classes regardless of their position within a word, and between all posttonic, pretonic and plain syllables classes (again, regardless of position), will be very high. The average similarity value will therefore also be high. If τ is set in relation to this high average, not all unstressed and all stressed syllable classes, respectively, will be merged, because not all syllable classes can exceed average similarity.

Average similarity is a language-specific property, and so is the number of syllable classes similar enough to be merged for a given τ value. For different generalised instantiations of the same prototype model to be comparable, object-set generalisation must have been carried out for each of them with the same τ value.

The threshold τ is best regarded as a variable parameter to the OS-generalisation procedure that can be used to control the degree to which a generalised OFS model will fit the data: the higher τ, the more closely the model will fit the data, and the less it will generalise over it. This is particularly appropriate in phonotactic modelling, because phonotactics seeks to encode not just the set of attested words, but also

Prototype OFS Model $\mathit{Syllable} = (\{\mathit{Syllable}, \mathit{Onset}, \mathit{Peak}, \mathit{Coda}\}, T, P, 2)$			
1:	*Syllable*	$\Rightarrow$	*Onset Peak Coda*
0:	*Onset*	$\Rightarrow$	$\{x \mid xay \in D, x \in \mathit{CONSONANTS}^*, a \in \mathit{VOWELS}\}$
	Peak	$\Rightarrow$	$\{x \mid yxz \in D, x \in \mathit{VOWELS}^+, y, z \in \mathit{CONSONANTS}^*\}$
	Coda	$\Rightarrow$	$\{x \mid yax \in D, x \in \mathit{CONSONANTS}^*, a \in \mathit{VOWELS}\}$

Figure 2: Simple prototype OFS model for syllable-level phonotactics.

æz, æʃ, ɑːsk, æsp, æs, æt, ɛt, ɔːk, ɔːks, ɑːnts, ɔː, ɔːz, æks, aɪ, aɪz, beɪ, bɑː, bɑːz, beɪb, bæk, bæks, siː, kæb, ʧɛə∗, ʧɛəd, sɪnʧ, sɪnʧt, kliːv, dɛf, diːl, djuːst, dʌvz, drɑːfts, dwɛld, faɪ, frɛt, gəʊld, grɒt, kwɪd, splæt, sprɪŋ, stræps, stʌn

Figure 3: Small data set of English monosyllabic words.

OFS Model $\mathit{Syllable} = (\{\mathit{Syllable}, \mathit{Onset}, \mathit{Peak}, \mathit{Coda}\}, T, P, 2)$			
1:	*Syllable*	$\Rightarrow$	*Onset Peak Coda*
0:	*Onset*	$\Rightarrow$	{ ϵ, b, s, k, st, f, d, ʧ, kl, dj, dr, dw, fr, g, gr, kw, spl, spr, str }
	Peak	$\Rightarrow$	{æ, ɑː, ɛ, ɔː, aɪ, eɪ, iː, ɛə, ʌ, əʊ, ɒ, ɪ, uː }
	Coda	$\Rightarrow$	{ ϵ, b, s, k, st, f, d, z, ʃ, sk, sp, ks, nts, ∗, nʧ, nʧt, v, l, vz, fts, ld, t, ŋ, ps, n }

Figure 4: Syllable-level phonotactic OFS model instantiated with set of English monosyllables.

OFS Model $\mathit{Syllable} = (\{\mathit{Syllable}, \mathit{Onset\_Coda}, \mathit{Peak}, \}, T, P, 2)$			
1:	*Syllable*	$\Rightarrow$	*Onset_Coda Peak Onset_Coda*
0:	*Onset_Coda*	$\Rightarrow$	{ ϵ, b, s, k, st, f, d, ʧ, kl, dj, dr, dw, fr, g, gr, kw, spl, spr, str, z, ʃ, sk, sp, ks, nts, ∗, nʧ, nʧt, v, l, vz, fts, ld, t, ŋ, ps, n }
	Peak	$\Rightarrow$	{æ, ɑː, ɛ, ɔː, aɪ, eɪ, iː, ɛə, ʌ, əʊ, ɒ, ɪ, uː }

Figure 5: OFS model of Figure 4 generalised with $\tau \leq 0.19$.

unattested, but wellformed words (often called 'accidental' gaps), while excluding only illformed words (or 'systematic' gaps). There is no objective dividing line between idiosyncratic and systematic gaps, and setting τ can be used as one way of controlling the degree of conservativeness in generalising over the set of attested words.

3.2 Example

As an illustration, consider the following example construction of a simple OFS model for syllable-level phonotactics (the constraints that hold on the possible phoneme sequences within syllables)[5]. The prototype OFS model constructed in the first step (Figure 2) encodes the standard assumption that the syllable-level phonotactics in different languages can be appropriately modelled by interpreting syllables as a sequence of consonantal phonemes (onset), followed by a sequence of vocalic phonemes (peak), and another sequence of consonantal phonemes (coda).

In the second construction step, a data set of English monosyllabic words (Figure 3) is used to instantiate the prototype OFS model. The instantiation procedure constructs an OFS model with new level 0 RHSs as shown in Figure 4. During OS-generalisation, *sim* values are computed for each pair of level 0 object sets. The only pairwise intersection that is non-empty (hence the only non-zero *sim* value) in this example is that between the sets *Coda* and *Onset* ($\mathit{sim} = 0.19$), which are merged if OS-generalisation is applied to OFS model *Syllable* with $\tau \leq 0.19$, resulting in the simpler, more general OFS model shown in Figure 5.

3.3 Language-Independent Prototyping

Language-independent prototyping (LIP) as a general approach to linguistic description seeks to define generic models that restrict — in some linguistically meaningful way — the set of grammars or descriptions that can be inferred from data. OFS modelling can be used as an implementational tool for LIP. Language-independent prototype OFS models can be defined by specifying a maximal number of objects and corresponding production rules such that when the prototype is instantiated and generalised with data sets from individual languages, dif-

---

[5] The example model is not intended to be a realistic phonotactic model, but is provided here merely as an illustration of the techniques outlined above.

$$\begin{array}{llll}
\hline
\multicolumn{4}{l}{\text{Prototype OFS Model } Word = (N, M, P, 2)} \\
\hline
1: & Word & \Rightarrow & S\_mon\_st\ + \\
 & & & S\_mon\_pl\ + \\
 & & & (S\_ini\_st\ S\_fin\_po)\ + \\
 & & & (S\_ini\_st\ S\_med\_po\ S\_med\_pl^*\ S\_fin\_pl)\ + \\
 & & & (S\_ini\_pr\ S\_fin\_st)\ + \\
 & & & (S\_ini\_pr\ S\_med\_st\ S\_fin\_po)\ + \\
 & & & (S\_ini\_pr\ S\_med\_st\ S\_med\_po\ S\_med\_pl^*\ S\_fin\_pl)\ + \\
 & & & (S\_ini\_pl\ S\_med\_pl^*\ S\_med\_pr\ S\_fin\_st)\ + \\
 & & & (S\_ini\_pl\ S\_med\_pl^*\ S\_med\_pr\ S\_med\_st\ S\_fin\_po)\ + \\
 & & & (S\_ini\_pl\ S\_med\_pl^*\ S\_med\_pr\ S\_med\_st\ S\_med\_po\ S\_med\_pl^*\ S\_fin\_pl) \\
0: & S\_mon\_st & \Rightarrow & \{x : {}'x \in D, x \in (M\backslash\{-\})^*\} \\
 & S\_mon\_pl & \Rightarrow & \{x : x \in D, x \in (M\backslash\{-,'\})^*\} \\
 & S\_ini\_st & \Rightarrow & \{x : {}'x - w \in D, x \in (M\backslash\{-\})^*\} \\
 & S\_ini\_pr & \Rightarrow & \{x : x -' vw \in D, x, v \in (M\backslash\{-\})^*\} \\
 & S\_ini\_pl & \Rightarrow & \{x : x - u -' vw \in D, x, v \in (M\backslash\{-\})^*\} \\
 & S\_med\_st & \Rightarrow & \{x : v -' x - w \in D, x \in (M\backslash\{-\})^*\} \\
 & S\_med\_pr & \Rightarrow & \{x : u - x -' vw \in D, x, v \in (M\backslash\{-\})^*\} \\
 & S\_med\_po & \Rightarrow & \{x : u'v - x - w \in D, x, v \in (M\backslash\{-\})^*\} \\
 & S\_med\_pl & \Rightarrow & \{x : (u'y - v - x - w \in D) \vee\ (u - x - v -' w \in D), x \in (M\backslash\{-\})^*\} \\
 & S\_fin\_st & \Rightarrow & \{x : w -' x \in D, x \in (M\backslash\{-\})^*\} \\
 & S\_fin\_po & \Rightarrow & \{x : w'v - x \in D, x, v \in (M\backslash\{-\})^*\} \\
 & S\_fin\_pl & \Rightarrow & \{x : w'v - u - x \in D, x, v \in (M\backslash\{-\})^*\} \\
\hline
\end{array}$$

Figure 6: Prototype OFS model for multi-syllable word-level phonotactics.

ferent object sets will be deleted and merged for different languages, resulting in different final, instantiated and generalised OFS models. In the following section, a language-independent phonotactic prototype OFS model is instantiated to surprisingly different OFS models for three closely related languages.

4 Multi-Syllable Phonotactic Models for German, English and Dutch

When applied to modelling multi-syllable word-level phonotactics, LIP with OFS Modelling means defining the maximum possible number of syllable classes that may be subject to different phonotactic constraints in a given group of languages. The exact set of syllable classes depends on the group of languages the prototype is intended to cover as well as the desired amount of generalisation over data (in general, a model that distinguishes only two syllable classes will generalise more than a model that distinguishes three or more classes, given the same data). The prototype presented in this section is intended to cover German, English and Dutch, and takes into account only phonological factors (syntactic factors such as word category which can also affect phonotactics are not taken into account). Two phonological factors are modelled: position of a syllable within a word, and position of a syllable relative to primary word stress.

For this modelling task, the LIP approach is implemented by constructing an OFS prototype model in which syllable classes reflecting all possible different combinations of position within a word and relative to stress are defined as level 0 uninstantiated object rules, and all possible ways in which the corresponding objects can be combined to form words are defined as higher-level object rules. No prior assumptions about where phonotactic variation occurs is hardwired into the model. Instead, the maximal ways in which phonotactics may vary in a group of languages is encoded. The idea is that prototype instantiation and OS-generalisation with data sets of phonological words from different languages will result in different final, instantiated phonotactic models.

4.1 Language-Independent Prototype OFS Model for Multi-syllable Phonotactics

The prototype model shown in Figure 6 distinguishes between twelve syllable classes which correspond to all possible combinations of position within a word and position relative to primary stress ($'$ marks primary stress, $-$ is the syllable separator, and S = syllable). As before, the set of all syllables is divided into four classes on the basis of position (*mon* = monosyllabic, *ini* = initial, *med* = medial, *fin* = final), each of which is divided further into four subclasses on the basis of stress (*st* = stressed, *pr* = pretonic, *po* = posttonic, *pl* = plain). This results in a total of 12 possible syllable cat-

	German		English		Dutch	
	all	unique (%)	all	unique (%)	all	unique (%)
Set_mon_st	5,028	849 (16.89%)	7,254	2,958 (40.77%)	5,641	719 (12.75%)
Set_mon_pl	1,813	1 (0.06%)	11	5 (45.45%)	0	- (-)
Set_ini_st	3,658	527 (14.41%)	3,345	409 (12.23%)	5,258	772 (14.68%)
Set_ini_pr	707	18 (2.55%)	2,560	1,328 (51.88%)	1,346	49 (3.64%)
Set_ini_pl	1,628	19 (1.17%)	1,495	437 (29.23%)	1,252	28 (2.24%)
Set_med_st	2,527	92 (3.64%)	1,600	90 (5.63%)	3,907	282 (7.22%)
Set_med_pr	618	12 (1.94%)	916	30 (3.28%)	1,026	26 (2.53%)
Set_med_po	2,518	66 (2.62%)	1,415	65 (4.59%)	3,296	185 (5.61%)
Set_med_pl	2,220	28 (1.26%)	1,156	82 (7.09%)	2,897	36 (1.24%)
Set_fin_st	4,261	822 (19.29%)	3,376	583 (17.27%)	4,972	803 (16.15%)
Set_fin_po	4,354	413 (9.49%)	4,141	882 (21.3%)	4,525	460 (1.02%)
Set_fin_pl	3,716	166 (4.47%)	2,635	306 (11.61%)	3,820	101 (2.64%)
TOTAL	10,598	3,013 (28.42%)	14,333	7,175 (50.06%)	11,443	3,461 (30.25%)

Table 4: Sizes of Level 0 object sets resulting from instantiations, and syllables unique to each set.

egories[6]. D is the data set given in instantiation, and M the corresponding set of terminals (here, the phonemic symbols that occur in D). The RHS of the level 1 object rule encodes all possible ways in which the twelve syllable classes can theoretically combine to form words. The prototype model is language-independent, because not all syllable classes will exist in all languages (e.g. a language where primary stress is always on the first syllable would not have classes of word-initial pretonic or plain syllables), and OS-generalisation will create different new syllable classes, depending on which classes are most similar in a given language.

4.2 Prototype Model Instantiations

Table 4 shows the sizes of the different level 0 object sets resulting from OFS model instantiations to the German, English and Dutch word sets derived from CELEX (the syllable sets are far too large to be shown in their entirety). In all three languages, the largest syllable set is the set of stressed monosyllables, and the smallest is the set of medial pretonic syllables[7]. Table 4 also shows (in the same format as in Section 2) the number of syllables in each syllable class that do not occur in any of the other classes.

In German and Dutch, percentages of unique syllables are significantly lower than in the classes reflecting position only and stress only that were shown in Section 2, indicating that some of the classes may not be worth distinguishing in phonotactic models. In English, however, the higher percentages of unique syllables are not far behind those shown previously, indicating that most of the twelve syllable classes in the prototype are worth distinguishing.

Some correlation is evident between the size of a set and the percentage of unique syllables it contains. In German, average syllable set size is $2,754$ and the average percentage of unique syllables is 6.48%. Five syllable sets are of above average size, and four of these also have above-average percentages of unique syllables. Seven syllable sets are below average in size, and non of these have above-average percentages of unique syllables. In English, the picture is not as straightforward. Average syllable set size is $2,717$, and average percentage of unique syllables is 18.62%. Of the four sets of above-average size, two have above-average, and two have below-average, percentages of unique syllables. Of the seven English syllable sets of below-average size (the set of plain monosyllables is disregarded again for English and Dutch), two have above-average, and five have below-average percentages of unique syllables. Finally, in Dutch, average set size is $3,449$ and average percentage of unique syllables is 6.33%. Four of the six above-average sized sets also have above-average percentages of unique syllables, while all of the below-average sized sets also have below-average percentages of unique syllables. However, there is no complete correlation, with some of the largest sets having very small percentages of unique syllables, and vice versa.

4.3 OS-Generalisation of Models

As is clear from the instantiation results presented in the preceding section, some syllable classes contain such low percentages of unique syllables that it is not worth distinguishing them as a separate class. OS-generalisation of models can be used to merge the most similar classes and reduce the number of classes that the model distinguishes.

---

[6] Not $4 \times 4 = 16$ classes, because some classes cannot exist (e.g. there is no such thing as a posttonic initial syllable).

[7] Disregarding the set of plain monosyllables of which there were no examples in the Dutch section of CELEX, and only a very small number in the English section.

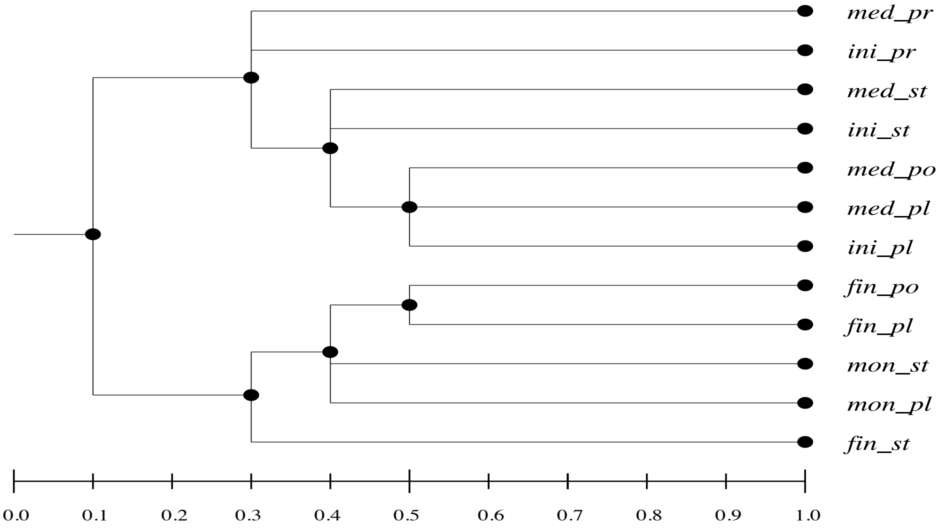

Figure 7: Cluster tree for German syllable sets.

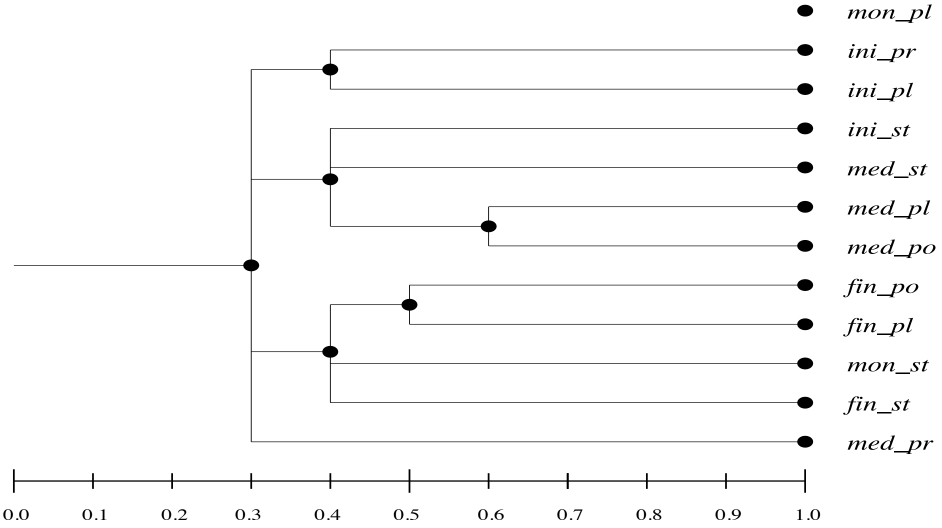

Figure 8: Cluster tree for Dutch syllable sets.

4.3.1 Generalisation of Multi-Syllable OFS Model for German

Figure 7 shows the cluster tree for the German syllable sets produced by carrying out OS-generalisation for $\tau = 0.1..1.0$ in increments of 0.1. Each node in the tree shows at which τ values the original syllable sets at the leaves dominated by the node were merged. The tree reveals a very neat picture for German. 0.56 is the highest τ value between any syllable class pair, so for $\tau \geq 0.6$ no classes are merged. $\tau = 0.5$ results in two clusters, one containing final unstressed syllables, the other initial and medial unstressed syllables. At $\tau = 0.4$, all monosyllables are added to the final syllable class, and one more medial and one more initial class to the set of initial and medial syllables. At $\tau = 0.3$, all monosyllables and final syllables on the one hand, and all initial and medial syllables on the other, are merged. Setting τ lower makes no difference until it is set below 0.2, at which point all of the original syllable classes are merged into a single set.

This shows clearly that in German the distinction between monosyllables and final syllables on the one hand, and between initial and medial syllables on the other, is very strongly marked (preserved even when τ is set as low as 0.2). This distinction is thus marked far more strongly than the unstressed/stressed division (which is more commonly encoded in DFA models of German phonotactics), which disappears at $\tau = 0.4$ (in fact, even earlier, at $\tau = 0.47$).

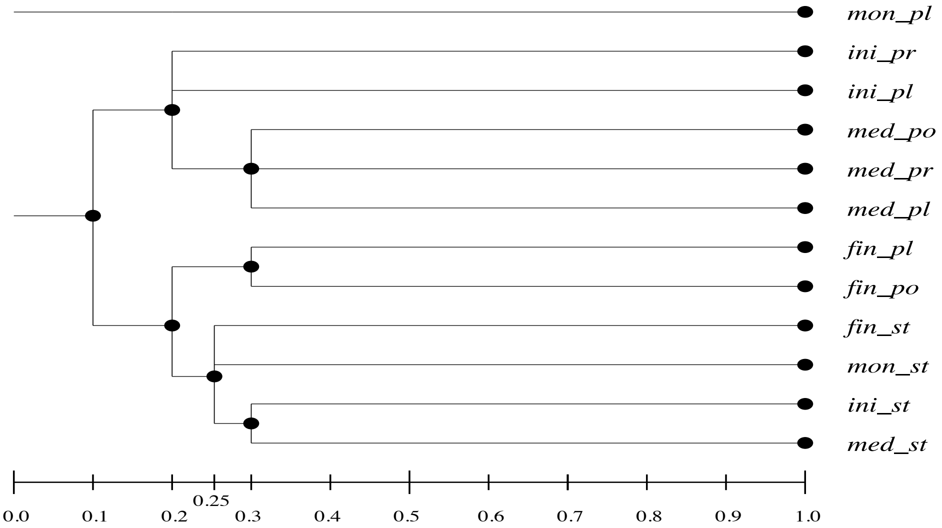

Figure 9: Cluster tree for English syllable sets.

4.3.2 Generalisation of Multi-Syllable OFS Model for Dutch

The cluster tree for Dutch (Figure 8) also reveals an important division between final and monosyllables on the one hand, and initial and medial syllables on the other. However, it is not as clearly marked as in German. There is a point ($\tau = 0.4$) when all final and monosyllables are in the same cluster, but this is not the case for the initial and medial syllables, which form subclusters that are correlated with stress. The medial plain and posttonic syllable sets are merged with each other at $\tau = 0.6$, and with the initial stressed and medial stressed syllables at $\tau = 0.4$. But there is no greater similarity between this cluster and the cluster of inital pretonic and plain syllables (formed at $\tau = 0.4$) than there is between it and the cluster of final and monosyllables. All three are merged into a single cluster at $\tau = 0.3$.

4.3.3 Generalisation of Multi-Syllable OFS Model for English

In the cluster tree for English (Figure 9), there are clusters clearly correlated with stress and clusters clearly correlated with position. At $\tau = 0.3$ three clusters are formed, one containing all medial syllable sets except the stressed medial syllables, another containing all final syllable sets except the stressed final syllables, and the third containing two stressed syllable sets. At $\tau = 0.25$, all stressed syllables together form one cluster. However, at $\tau = 0.2$, two unstressed syllable sets are added to this cluster, while all the remaining unstressed sets form the other large cluster. Thus, in English, both stress and position are strong determinants of phonotactic variation, but differences resulting from stress are more pronounced than those resulting from position.

4.4 Discussion

The LIP approach implemented with OFS Modelling proceeds in three steps. First, the factors likely to produce phonotactic idiosyncracy (stress and position in the above examples), and the constituents to be used in the analysis (syllables only in the above examples), are decided, and a prototype model is constructed on this basis. This prototype distinguishes as many objects at level 0 as there are possible combinations of factors and lowest-level constituents. All ways in which these objects can combine to form higher-level constituents are encoded at the corresponding higher levels in the model.

In the second step, the prototype is instantiated with data sets from different languages. The degree to which the instantiated models generalise over the given data is determined by the number of constituents and subcategories of constituents distinguished in the prototype. As an example, consider the different degrees to which three models that discriminate different numbers of syllable classes generalise over given data. All three models define words as sequences of syllables, and syllables as sequences of phonemes. The first model has only one syllable class, the second distinguishes four classes reflecting position in a word, and the third is the same as the model presented in the preceding section, i.e. distinguishes twelve syllable classes. After instantiation with the same data set of German phonological word forms from CELEX used previously, the three models will encode supersets of the data set that generalise over it to different degrees. Looking at subsets of words of the same length gives some impression of the differences. For instance, model 1 encodes 10,598 monosyllabic German words (the total number of different syllables in the data), whereas models 2 and 3 encode only $6,841$ monosyllables (the actual number of monosyllabic words in CELEX). The following table shows the number of bisyllabic words each model encodes.

Model	Bisyllabic words
(1) *Syll Syll*	1.12×10^8
(2) *Syll_ini Syll_fin*	2.67×10^7
(3) (*Syll_ini_pr Syll_fin_st*)+ (*Syll_ini_st Syll_fin_po*)	1.89×10^7
Attested forms	7.09×10^4

Model 3 permits about 266 times as many bisyllabic word forms as there are in CELEX, model 2 encodes 1.4 times as many as model 3, and model 1 encodes 4.2 times as many as model 2. Thus, through progressively finer grained subcategories of syllables, progressively closer approximations of the set of attested forms can be achieved.

However, doing this in an indiscriminate, language-independent way may produce some syllable classes that are very similar. With OS-generalisation, the most similar classes can be merged, so that only strongly marked differences are preserved. However, setting τ to any specific value is problematic. Producing cluster trees with a range of τ values can give some idea of important class distinctions, and can be used as a basis for determining an appropriate τ value. τ can further be motivated by different linguistic assumptions and the intended purpose of the generalised models. Generalising different instantiations of the same prototype for the same τ value, makes it possible to compare the relative markedness of phonotactic variation in different languages.

5 Summary and Further Research

This paper described how OFS modelling and the multi-syllable approach can be combined with language-independent prototyping to create a method for designing phonotactic models that (i) facilitates automatic model construction, (ii) produces models that are arbitrarily close approximations of the set of wellformed phonological words in a given language, and (iii) provides a generalisation method with control over the degree to which final models fit given data. Extensions of the approach currently under investigation include stochastic OFS models, and the integration of OFS models into finite-state syntactic grammars.

References

R. H. Baayen, R. Piepenbrock, and L. Gulikers, editors. 1995. *The CELEX Lexical Database (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Anja Belz. 1998. An approach to the automatic acquisition of phonotactic constraints. In T. Mark Ellison, editor, *Proceedings of SIGPHON '98: The Computation of Phonological Contraints*, pages 35–44.

Anja Belz. 2000. *Computational Learning of Finite State Models for Natural Language Processing.* Ph.D. thesis, School of Cognitive and Computing Sciences, University of Sussex.

S. Bird, editor. 1991. *Declarative Perspectives on Phonology*, volume 7 of *Edinburgh Working Papers in Cognitive Science.* Centre for Cognitive Science, University of Edinburgh.

Julie Carson-Berndsen. 1992. An event-based phonotactics for German. Technical Report ASL-TR-29-92/UBI, Fakultät für Linguistik und Literaturwissenschaft, University of Bielefeld.

Julie Carson-Berndsen. 2000. Finite state models, event logics and statistics in speech recognition. In *Philosophical Transactions of the Royal Society*, volume A 358. Royal Society. In press.

J. S. Coleman and J. Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology, SIGPHON '97*, pages 49–56.

John A. Goldsmith. 1990. *Autosegmental and Metrical Phonology.* Blackwell, Cambridge.

Jusek, Fink, Kummert, Sagerer, Berndsen, and Gibbon. 1994. Detektion unbekannter Wörter mit Hilfe phonotaktischer Modelle. In W. Kropatsch and H. Bischof, editors, *Mustererkennung '94, 16. DAGM-Symposium und 18. Workshop der ÖAGM Wien*, pages 238–245.

A. Jusek, A. Fink, F. Kummert, and G. Sagerer. 1996. Automatically generated models for unknown words. In *Proceedings of the Sixth Australian International Conference on Speech Science and Technology*, pages 301–306.

Michael Mastroianni and Bob Carpenter. 1994. Constraint-based morpho-phonology. In *Proceedings of the First Meeting of the ACL Special Interest Group in Computational Phonology, SIGPHON '94.*

Marc A. Zissman. 1995. Language identification using phoneme recognition and phonotactic language modelling. In *Proceedings of ICASSP '95*, volume 5, pages 3503–3506.