Multi-level OT

An argument from speech pathology

Dirk-Bart den Ouden
University of Groningen

1. Introduction

Since the early 1990s, Optimality Theory (OT) has quickly gained ground in phonology. Its main appeal lies in two characteristics: the focus on well-formedness of the output and the softness of constraints, where a constraint can be violated in order to satisfy more important requirements.

A conspicuous characteristic of classic OT is that all constraints on the output should compete with each other at all times. This basically means that the 'construction' of the output occurs in one step. OT is therefore minimally derivational, the only derivation being that from the input to the output. Mainly on the basis of data from aphasic speakers, I argue that the 'one-step' variant of OT lacks psychological validity and that it is better to assume that the OT algorithm plays a role in determining (phonological) structure at different cognitive levels of processing, at least in production.

2. Single-level OT

One of the criticisms of OT has been based on phonological processes of which it is argued that they simply cannot be adequately described without making reference to some notion of cyclicity or multiple levels of processing.

Such criticism has led to a number of adaptations to the original theory (Prince & Smolensky 1993), all aimed at giving satisfactory descriptions of morphophonological processes in which the output form seems to be opaque, and in which certain constraints appear to have been applied only to specific substrings of the eventual output form. Examples of such tools that aim to

Linguistics in the Netherlands 2004 21 (2004), **146–157.**
ISSN 0929–7332 / E-ISSN 1569–9919

maintain the one-step evaluation are Output-Output Correspondence (McCarthy and Prince 1995), in which the optimal output form wants to be as similar as possible to other output forms it is related to, and Sympathy Theory (McCarthy 1998), in which the optimal output form wants to resemble a fairly arbitrarily chosen other output candidate.

Other optimality theorists have chosen to abandon the one-step derivation altogether and to incorporate some type of sequentiality in OT, allowing multiple levels of evaluation, with constraints that apply only to specific stages in the derivation (Booij 1997; Rubach 2000). Crucially, the 'founding fathers' of OT, Prince and Smolensky (1993:79), did not put an absolute restriction on the theory as having only one level of evaluation, although the current practice is such that multiple levels of evaluation are considered a weakness.

3. Multi-level cognitive processing

Beside this formal discussion stands a large body of evidence for multiple levels of processing, from the fields of psycho- and neurolinguistics (e.g. Levelt 1989). Imaging studies of brain activity during language processing, as well as lesion studies, show that different parts of the brain are involved in the performance of different language functions (Démonet 1998; Whitaker 1998). Results of studies into temporally successive brain activity point towards a 'phonological loop' (Baddeley 1986) in which abstract and articulatory levels are distinct, though possibly mutually influential.

Of course, a formal theory of grammar, such as OT, is not *designed* for the purpose of reflecting the psychological reality of language performance. It is specifically meant to reflect competence, so one might argue that any knowledge (however limited) about what actually goes on during speech production and the building of language structure is irrelevant to the formal grammar. However, OT does rely on support from sources external to the grammar, such as language acquisition data and arguments of learnability. It will be hard to maintain that these factors are not rooted in the reality of language use. As such, OT does make claims to psychological reality, which means it is bound by logic to all relevant aspects of this reality.

4. The coda-observation in fluent and nonfluent aphasia

4.1 Experiment

Ten fluent and ten nonfluent Dutch aphasic speakers were tested with a monosyllabic repetition task, on order to investigate the influence of positional

markedness within syllables on their paraphasic output (Den Ouden 2002). From language acquisition data, language change and typology, it is known that certain syllable positions are more prone to error than others. For example, in onset clusters, sonorant consonants will be deleted sooner than obstruents. The question was whether this pattern might be related to a phonological or a phonetic level of speech output (planning).

In the absence of extralinguistic factors, such as dysarthria, nonfluent patients with distorted phonological output are claimed to suffer from difficulty in the timing and co-ordination of articulatory movements (Blumstein et al. 1980). This is related to a deficit at a cognitive phonetic level of processing (Code 1998), peripheral to the language processing system, but still considered linguistic. These nonfluent speakers have apraxia of speech.

Fluent aphasic patients presenting with literal paraphasias have unimpaired articulation, but suffer from a deficit in the appropriate selection of phonemes. The label of fluent aphasia covers a range of traditional syndromes, such as Wernicke's aphasia and conduction aphasia. What these disorders have in common is that they yield incorrect phonological plans. This may be caused by incorrect lexical access or representations, or by incorrect phonemic sequencing, i.e. the mapping of speech sounds and features onto metrical frames (phonological encoding). The difference, then, between fluent and nonfluent aphasic speakers is that fluent aphasics create an erroneous phonological plan that may be correctly executed phonetically, whereas nonfluent aphasics phonetically implement incorrectly a correct phonological speech plan.

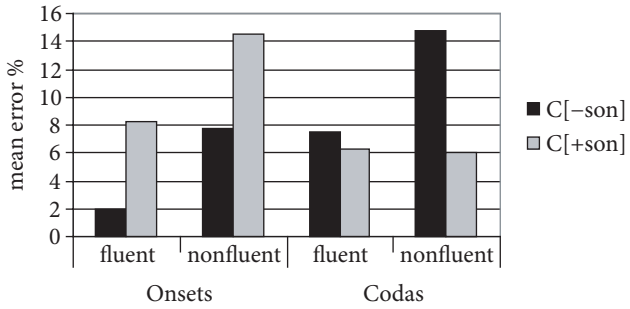
The subjects repeated 114 Dutch monosyllabic words. Their segment deletions were counted per syllable position, defined on the basis of a template. Following this template, in example word ***sprints*** the boldface positions (p, i and t) are strongest, i.e. least prone to deletion. Results showed that the deletion patterns of fluent and nonfluent speakers were largely similar, with the positions stipulated as relatively weak in the template indeed being deleted more often than their stronger neighbours. However, in coda clusters, the pattern was not so clear, which is why a further analysis was done, specifically aimed at clusters.

4.2 The coda observation

For this analysis, only 41 items from the original task were analysed. These were the items with complex onsets or complex codas which did not violate the sonority slope, meaning that the sonority value of segments rose from the margins to the peak. The clusters all consisted of one obstruent (C[−son]) and one sonorant (C[+son]) segment.

Results are visualised in figure (1), where the black bars show the mean (proportionate) number of deletions of obstruents and the grey bars those of the sonorants.

(1) Deletions in onset and coda clusters for fluent and nonfluent speakers



The patterns of deletions are equal for both groups in onsets, but not in codas, where only the nonfluent patients delete the sonorant coda position significantly more often than the nonsonorant coda position. For example, the nonfluent patients' rendition of the target word *print* /prɪnt/ will characteristically be [pɪt], while the fluent patients, as a group, will turn it into either [pɪn] or [pɪt], in a seemingly random fashion.

Apparently, the positional markedness relations yield similar output patterns in the paraphasic syllable onsets produced by fluent and nonfluent aphasic speakers, but different patterns in codas. This 'coda observation' leads to a different analysis of the data, this time in terms of conflicting types of markedness, segmental vs. syllabic (or 'positional').

4.3 Segment vs. Syllable Markedness

According to a hierarchy of segmental markedness going back to Jakobson (1941), consonants are less marked if they are less sonorant. This segmental markedness applies non-contextually; it does not take into account the position of a segment within a syllable. Note that it can account for the error pattern of the nonfluent patients, but not for the coda observation in fluent patients' errors.

If we do look at segments in the context of prosodic structure, however, a different picture emerges. Clements (1990) argues that the preferred sonority slope of syllables has a steep rise in sonority pre-vocalically and a slow decline in sonority postvocalically. This Sonority Cycle allows us to formulate a second markedness hierarchy, viz. syllable markedness, according to which onsets want to be nonsonorant and codas want to be sonorant.

Syllable markedness alone cannot account for the pattern of deletions observed in fluent patients' paraphasias. However, segmental and syllable markedness reinforce each other in onsets, whereas they are in opposition in codas. It is this combination of the two types of markedness, or rather the crucial conflict between them, that may account for the fluent patients' error pattern, which is a 50–50 distribution of deletions in coda clusters.

The full analysis of the presented data is that nonfluent aphasics have a deficit at a phonetic level of processing, at which the markedness of individual segments, or feature combinations, is an influential factor. The impairment allows this type of markedness to become dominant, which means that when clusters of consonants are reduced, the nonsonorant, segmentally least marked consonant will come out as the winner, irrespective of its position within a syllable. But before this phonetic level of processing, constraints on sonority sequencing, i.e. on preferred syllable structure, are active beside constraints on segmental markedness. At the affected level(s) of processing in fluent aphasic patients, the conflict between segmental markedness constraints and syllable markedness constraints emerges, as structure preserving constraints lose control over the output of the speech production process. For fluent aphasics, this yields a pattern of errors in which onsets are relatively systematically reduced to nonsonorant segments, while codas are reduced on a seemingly random basis to either sonorant or nonsonorant segments, as the constraints are in direct conflict over what is a preferred segment in coda position.

5. An OT analysis

The tools of OT seem particularly suited to give a representation of the conflict between markedness constraints with opposing output goals, such as described above. However, the application of OT to language pathology is only in its infancy, so a number of choices have to be made and argued for in this domain.

5.1 Aphasia

Aphasia is generally characterised by a prominence of unmarked structures. Compared to speakers without brain damage, the aphasic speaker is less faithful to the input, the input here being the lexicon or, for example, utterances to be repeated. The most straightforward way of representing this in OT is by a lowering of faithfulness constraints, relative to markedness constraints.

Note that it is theoretically also possible that input representations themselves are disturbed, so that the correct constraint ranking works on incorrect input, or that the number and/or type of output candidates that are generated is in some way restricted. In OT, however, these options do not directly account for the prominence of the unmarked, as observed in aphasic speech. Any systematic way of constricting the input or the output candidates would somehow have to incorporate extra markedness constraints on these domains. This would come down to an extratheoretical add-on for which there is no evidence or argument in nonpathological natural language. I assume that language impairment is focal breakdown of the normal language system,

crucially within its own terms. The impairment will not to add new features to the normal system. The aphasic patients in this study show markedness effects in their impaired output, and the OT representation of the impairment consists of the lowering of faithfulness constraints, allowing markedness constraints to have greater influence on the selection of the optimal output candidate.

Aphasic data are never homogeneous. There is much noise and variability, which is precisely why statistics are used to determine whether some structures are really used more often than others, or whether differences might be due to chance. Variation, or optionality, can be represented by 'switching' of adjacent constraints (e.g. Tesar and Smolensky 1998). The *Gradual Learning Algorithm* of Boersma and Hayes (2001) is even able to capture statistical differences in the frequency of occurrence of certain forms. Constraints have moving ranges along the hierarchical scale, which are interpreted as probability distributions, i.e., they are normal distributions with the ranking value as their peak. These ranges may overlap. At one particular moment of evaluation, the position of a constraint on the ranking scale, i.e. its selection point, is less likely the further it is from this constraint's ranking value. Through this, one is able to calculate the probability of a certain ranking of constraints at the moment of constraint evaluation.

Of course, the output of aphasic speakers is not consistently erroneous either. Depending on the severity of their impairment, aphasic speakers will often produce the correct (target) output form. In these cases, it is assumed that the selection point of the relevant faithfulness constraint is above the relevant markedness constraints. The scope of FAITHFULNESS, therefore, should at least partially overlap with that of these markedness constraints. In the analysis provided in the following section, I will focus on the constraint rankings underlying paraphasic output.

5.2 Constraints and tableaux

The constraints required for the analysis presented here are given in (2).

(2) *Markedness*
*C[+SON] Do not allow sonorant consonants
HONS (Onset Harmony) No sonorant material in onsets
HCOD (Coda Harmony) No nonsonorant material in codas
Faithfulness
PARSE Preserve input material

As discussed above, segmental markedness seems to be prominent in the paraphasias of the nonfluent aphasic speakers, who have an impairment at the cognitive phonetic level of speech planning. This is represented in Tableau (3), where *C[+SON] is ranked higher than PARSE. The markedness constraint here picks out the only relevant candidate that does not violate it.

(3) Tableau nonfluent patients /prɪnt/ → /pɪt/

/prɪnt/	*C[+son]	Parse
prɪnt	** !	
pɪnt	* !	*
☞ pɪt		**
pɪn	* !	**
rɪt	* !	**
rɪn	** !	**
rɪnt	** !	*
prɪt	* !	*
prɪn	** !	*

The fluent patients turn example word *print* into either [pɪn] or [pɪt]. This is because of a competition between a constraint on the preferred sonority value of the syllable constituent coda and the segmental markedness constraint that disallows sonorant consonants. With respect to HCod and *C[+Son], two rankings are possible, with different results, as shown in Tableaux (4a) and (4b).

(4) a. Tableau fluent patients /prɪnt/ → /pɪt/

/prɪnt/	*C[+son]	HCod	HOns	Parse
prɪnt	** !	*	*	
pɪnt	* !	*		*
☞ pɪt		*		**
pɪn	* !			**
rɪt	* !	*	*	**
rɪn	** !		*	**
rɪnt	** !	*	*	*
prɪt	* !	*	*	*
prɪn	** !		*	*

b. Tableau fluent patients /prɪnt/ → /pɪn/

/prɪnt/	HCod	*C[+son]	HOns	Parse
prɪnt	* !	**	*	
pɪnt	* !	*		*
pɪt	* !			**
☞ pɪn		*		**
rɪt	* !	*	*	**
rɪn		** !	*	**
rɪnt	* !	**	*	*
prɪt	* !	*	*	*
prɪn		** !	*	*

Note that the specific ranking of HOns is irrelevant here, as it is not in conflict with the other markedness constraints. I have represented it in the tableaux, because HOns is linked strongly to HCod, both being similar types of constraints on syllable content. The reason for ranking it below HCod is that codas are generally more marked than onsets, so it seems reasonable to assume that restrictions on coda content are more important than restrictions on onsets.

5.3 Level-specific constraints vs. large-scale reranking

The data of the fluent aphasic speakers come about by the switching of positions between HCod and *C[+Son] and, of course, the lowering of Parse, with respect to its 'normal' position above these markedness constraints. Apparently, then, at the level of deficit of the fluent patients, HCod and *C[+Son] are so closely ranked that their ranges overlap almost 100%. What we could do now, is to say that at the level of deficit of the nonfluent patients, the ranking of these constraints is very different, i.e., their ranking values are much further apart, so that *C[+Son] is always most prominent. This is an undesired situation, as it opens up the possibility of totally different rankings at (or: in the representation of) different levels of processing. A major argument against an analysis in which different types of aphasia are represented through structural reranking of markedness constraints is the fact that aphasic speech errors hardly ever violate the phonotactics of the mother tongue of the speaker, or, indeed, universal restrictions on well-formedness (see Buckingham 1992). This would be unexplained if markedness constraints changed position in the hierarchy on a large scale. The adherence to (mother tongue) phonotactics points towards a lowering of faithfulness constraints only.

However, the variation found in the patterns of paraphasias belonging to different types of aphasia, such as observed in the presently discussed study, acts as an argument *against* the mere lowering of faithfulness constraints in the representation of aphasia. To represent different aphasic symptoms only through different degrees of faithfulness lowering comes down to saying that aphasic 'syndromes', or rather, clusters of symptoms, only differ with respect to the degree of severity of the impairment. This is a view that has indeed been held (e.g. Freud 1891), but detailed (linguistic) analysis of aphasic data lead contemporary aphasiologists to think of different types of aphasia as reflecting impairments at different functional levels of cognitive (if not linguistic) processing. This is also the approach adopted here. At those particular levels, of course, the impairments may still differ in degree of severity.

For these reasons, rather than claiming that the constraints HOns and HCod are ranked noncompetitively low at the level of impairment of nonfluent aphasics, I argue that they are *non-existent* at this level. In this way, structural reranking of markedness constraints is avoided. This means, then, that the analysis allows different levels of evaluation of constraints, where not all constraints are active (i.e. exist) at all levels. I maintain that, from a psycholinguistic and neurolinguistic perspective, this is the only natural way to conceive of linguistic processing.

In psycholinguistic modelling, it is common practice to minimise the number of levels, modules or stages of processing to those necessary for an accurate representation of empirical findings. A similar principle, Level Minimalism, is formulated by Rubach (2000) for his modification of OT, Derivational Optimality Theory, which allows multiple levels of evaluation. Another principle he formulates to restrict the power of his framework is that of Reranking Minimalism: "[the] number of rerankings is minimal [—] reranking of constraints comes at a cost and needs to be argued for" (Rubach 2000: 313). This principle is in line with the present approach of unstable rankings to account for variation and level-specific constraints instead of structural reranking of markedness constraints to account for the influence of different factors at different levels of speech production processing. The idea is that the only difference between the impaired and the healthy system is the lowering of faithfulness constraints at the affected level(s) of processing.

Another treat offered by this analysis is that syllables and constraints on their structure are only relevant as organising units at phonological levels of processing and not at the cognitive phonetic level of processing where articulation is planned. Syllables are not articulatory units (for a discussion of corroborative evidence for this claim, see Den Ouden 2002: 89–90).

5.4 OT for unimpaired language

According to the approach presented here, language impairment at specific levels of processing brings to light the factors that are functional at these levels. In OT, these factors are represented by constraints. Some of these constraints will normally, in an unimpaired language system, not be 'relevant', as they are hidden under a layer of faithfulness constraints, except in child language acquisition, when faithfulness constraints are also assumed to be ranked low. Thus, in non-brain-damaged speakers, the same constraints are functional at the same levels of processing as in aphasic speakers (of the type discussed here). However, for non-impaired speakers, faithfulness constraints are ranked sufficiently high at all these levels to ensure 'normal' native language output. Lowering of faithfulness, as in language impairment, causes the 'emergence of the unranked'.

This implies that it is still not *a priori* impossible to represent the grammar of unimpaired language in a single constraint tableau. Even if the process of speech production will work such that the output of one level, or module, serves as the input to the next, the static description of the language system can incorporate the various factors of influence in one representation. It is only when the system breaks down that the individual parts reveal themselves.

6. Conclusion

On the basis of fluent aphasics' and nonfluent aphasics' responses to a monosyllabic real word repetition test, I have argued that there is a difference between the phonological level of speech processing and the phonetic level of processing. In an OT approach, the constraints responsible for preferred syllable content are active only at pre-phonetic levels of evaluation, whereas a constraint on segmental markedness, saying that consonants should be as consonantal as possible (and therefore nonsonorant) is active at the pre-phonetic, as well as at the phonetic level.

Aphasia, in this approach, comprises the lowering of faithfulness constraints at the affected level of processing. The different levels account for the different types of aphasia that are generally recognised. Also, aphasia is characterised by highly variable output caused (or at least represented) by unstable ranking of close (adjacent) markedness constraints. In the approach presented here, the lowering of faithfulness constraints makes visible the unstable rankings that are there in the first place, but which are normally hidden because they do not have an effect on normal speech output.

Language breakdown provides a window on the workings of the language system and linguistic theories should be able to deal with the view thus offered.

It is not sufficient to claim that OT is 'not about' breakdown or psycholinguistic models, or that it does not have to be able to account for evidence of temporal processing or multiple levels of processing as long as there is no straightforward theory of the relation between language, mind and brain. Through such argumentation, phonological theory runs the danger of becoming merely a boundlessly creative method of deriving surface level data (output forms) from hypothesised underlying (input) forms. Therefore, OT should aim at ways to incorporate multiple levels of (phonological) processing, rather than focus on retaining the single evaluation hypothesis. The present paper has discussed one way of dealing with such stages in processing.

References

Baddeley, A. (1986) *Working Memory.* Clarendon Press, Oxford.

Blumstein, S. E., Cooper, W. E., Goodglass, H., Statlender, S. and Gottlieb, J. (1980) 'Production deficits in aphasia: A voice-onset time analysis'. *Brain and Language* 9, 153–70.

Boersma, P. and Hayes, B. (2001) 'Empirical tests of the gradual learning algorithm'. *Linguistic Inquiry* 32, 1, 45–86.

Booij, G. E. (1997) 'Non-derivational phonology meets Lexical Phonology'. In I. Roca, ed., *Derivations and Constraints in Phonology.* Oxford University Press, Oxford.

Buckingham, H. W. (1992) 'Phonological production deficits in conduction aphasia'. In S. E. Kohn, ed., *Conduction Aphasia.* Lawrence Erlbaum Associates, Hillsdale etc.

Clements, G. N. (1990) 'The role of sonority in core syllabification'. In J. Kingston and M. E. Beckman, eds., *Papers in Laboratory Phonology I. Between the Grammar and Physics of Speech.* Cambridge University Press, Cambridge.

Code, C. (1998) 'Models, theories and heuristics in apraxia of speech'. *Clinical Linguistics and Phonetics* 12, 1, 47–65.

Démonet, J. F. (1998) 'Tomographic Brain Imaging of Language functions: Prospects for a New Brain/Language Model'. In B. Stemmer and H. A. Whitaker, eds., *Handbook of Neurolinguistics.* Academic Press, San Diego, 131–142.

Freud, S. (1891) *Zur Auffassung der Aphasien: Eine kritische Studie.* Franz Deuticke, Leipzig and Vienna. In E. Stengel, transl., *On Aphasia: A Critical Study.* International Universities Press, New York, 1953.

Jakobson, R. (1941) *Kindersprache, Aphasie und allgemeine Lautgesetze.* Translation: A. R. Keiler (1968) *Child Language, Aphasia and Phonological Universals.* Mouton, The Hague.

Levelt, W. J. M. (1989) *Speaking: From Intention to Articulation.* MIT Press Cambridge, Massachusetts.

McCarthy, J. (1998) 'Sympathy and phonological opacity', Ms., University of Massachusetts, Amherst.

McCarthy, J. and Prince, A. (1995) 'Faithfulness and Reduplicative Identity'. In J. Beckman et al., eds., *Papers in Optimality Theory*, University of Massachusetts Occasional Papers 18, 249–384.

Ouden, D. B. Den (2002) *Phonology in Aphasia: Syllables and segments in level-specific deficits.* Doc. diss., University of Groningen.

Prince, A. and Smolensky, P. (1993) *Optimality Theory: Constraint interaction in generative grammar.* Rutgers University Center for Cognitive Science Technical Report 2.

Rubach, J. (2000) 'Glide and glottal stop insertion in Slavic languages: A DOT analysis'. *Linguistic Inquiry* 31, 2, 271–317.

Tesar, B. and Smolensky, P. (1998) 'Learnability in Optimality Theory'. *Linguistic Inquiry* 29, 229–268.

Whitaker, H. A. (1998) 'Neurolinguistics from the Middle Ages to the Pre-Modern Era: Historical vignettes'. In B. Stemmer and H. A. Whitaker, eds., *Handbook of Neurolinguistics.* Academic Press, San Diego, 27–54.