A NOTE ON PHONOLOGICAL SIMILARITY IN TESAR'S (2013) THEORY OF OUTPUT-DRIVENNESS

GIORGIO MAGRI

Abstract. Tesar's (2014) notion of *output-drivenness* is an attempt at characterizing non-opaque phonological mappings in terms of a notion of phonological similarity between underlying and surface levels. Tesar defines phonological similarity concretely in terms of segment strings and correspondence relations. Magri (2017) instead defines similarity axiomatically through an inequality on faithfulness constraint violations. This paper studies the formal properties of the latter axiomatic notion of similarity. In particular, it shows that Tesar's concrete definition of similarity satisfies Magri's axiomatization in terms of faithfulness constraints. The theory of output-drivenness reconstructed in Magri (2017) thus subsumes Tesar's (2014) original theory as a special case.

Tesar's (2014) *output-drivenness* is a formal condition on phonological grammars, construed as mappings from underlying to surface (or *output*) forms. It demands that a grammar which maps an underlying form UR_1 to some surface form SR also maps to that surface form SR any other underlying form UR_2 such that UR_2 is more similar to SR than UR_1 is. The notion of output-drivenness is relevant to phonological theory, because non-output-drivenness formally unifies various opaque phonological phenomena. Furthermore, Tesar shows that output-drivenness has learnability implications in the context of the classical *inconsistency detection* approach (Merchant 2008) to the problem of learning a lexicon of underlying forms from a paradigm of surface forms.

The notion of output-drivenness is predicated on a notion of phonological *similarity*: we need a way to compare a mapping $(\mathrm{UR}_1, \mathrm{SR})$ to another mapping $(\mathrm{UR}_2, \mathrm{SR})$ to establish that the underlying and surface forms in the latter mapping are more similar to each other than the underlying and surface forms in the former mapping. Tesar proposes a notion of phonological similarity *concretely* defined in terms of *disparities* between strings of segments and the correspondence relations (in the sense of McCarthy and Prince 1995) that hold among those segments. Being concretely defined in terms of the basic building blocks of phonological representations, Tesar's notion of similarity has the advantage of being framework-independent. The corresponding notion of output-drivenness thus bridges frameworks as different as rule-based and constraint-based phonology. This is important because the phonological relevance of (non)-output-drivenness lies in its ability to characterize opaque processes *extensionally* (namely, at the framework-independent level of mappings from underlying to surface forms) rather than *intensionally* (namely, in terms of framework-dependent notions such as counter-feeding and counter-bleeding rule orderings).

In Magri (2017), I explore an alternative approach whereby phonological similarity is defined not at the level of strings and correspondence relations but through an axiom on faithfulness constraint violations. Despite its reliance on faithfulness constraints, I submit that my alternative approach is not inconsistent with the desired framework-independency of output-drivenness. To start, the technical notion of *disparity* that Tesar relies on for its definition of similarity is really just a different name for a faithfulness constraint violation. Furthermore, Tesar does not shy away from correspondence relations despite the fact that they are a representational device needed

Date: December 19, 2017.

This paper has benefitted greatly from discussion with Bruce Tesar. The material in section A.5 has been conceived jointly in a dark night at Rutgers University. The research reported in this paper has been funded in part by a grant from the *Agence Nationale de la Recherche* (grant number ANR-16-CE38-0004-01).

by constraint-based phonology to get around the lack of phonological derivations. In fact, Tesar (p. 34) explains that, "while in linguistics the terminology of correspondence is perhaps found most explicitly in the OT literature, the concept is equally important to any generative theory. There is a correspondence relation implicit in every SPE-style rule." Crucially, the same argument applies to faithfulness constraints: although they were only formalized in OT, faithfulness considerations are plausibly intrinsic to phonological theorizing, independently of the framework. Finally, Tesar's concrete definition of similarity is tailored to restrictive representational assumptions, which only allow for deletion, epenthesis and feature mismatches but no additional faithfulness violations. Indeed, Tesar's application of output-drivenness to OT presupposes a faithfulness constraint set limited to MAX, DEP and IDENT constraints. An axiomatization of similarity in terms of an *arbitrary* set of faithfulness constraints allows Tesar's theory of output-drivenness to be generalized to richer representational assumptions, as shown in Magri (2017).

This paper thus looks closer at the formal properties of Magri's (2017) alternative notion of phonological similarity and in particular at its relationship with Tesar's (2014) original notion of similarity. Section 1 sets the background: it recalls Tesar's (2014) notion of output-drivenness, it reviews Tesar's concrete definition of similarity, and it complements it with Magri's (2017) alternative definition based on an axiom on the faithfulness constraints. Section 2 shows that Tesar's concrete definition of similarity is a special case of the axiomatic definition. This relationship between the two notions of similarity can be made to follow from an intermediate result in the sophisticated analysis developed in chapter 3 of Tesar's book. For completeness, the final appendix provides a self-contained proof. This proof largely consists of rebooting various ingredients of Tesar's analysis, apart for a "trick" which allows the analysis of MAX to be completely reduced to the analysis of DEP, yielding a substantial simplification. Section 3 looks at the formal properties of phonological similarity and establishes conditions on the candidate set and the faithfulness constraint set which ensure that the axiomatic definition of similarity proposed in Magri's (2017) yields a relation among pairs (UR, SR) of underlying and surface forms which qualifies as a partial order. I conclude in section 4 that Magri's (2017) definition of similarity provides a suitable axiomatization of Tesar's (2014) concrete definition of similarity and that the theory of output-drivenness reconstructed in Magri (2017) therefore subsumes Tesar's (2014) original theory as a special case.

1. BACKGROUND: OUTPUT-DRIVENNESS AND PHONOLOGICAL SIMILARITY

This section recalls Tesar's (2014) notion of output-drivenness and details two approaches to the definition of phonological similarity that output-drivenness is predicated on: Tesar's concrete definition and Magri's (2017) definition in terms of an axiom on the faithfulness constraints.

1.1. **Output-drivenness**

Tesar assumes the representational framework (1), which is a segmental version of McCarthy and Prince's (1995) *Correspondence Theory.* Underlying and surface forms are strings of segments. By (1a), phonological candidates establish a correspondence between their segments. Segments are denoted by $\mathsf{a}, \mathsf{b}, \mathsf{c}, \ldots$ and strings by $\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$ The notation $\mathbf{a} = \mathsf{a}_1 \cdots \mathsf{a}_\ell$ says that the string $\mathbf{a}$ is the concatenation of the segments $\mathsf{a}_1, \ldots, \mathsf{a}_\ell$. By (1b), a phonological grammar effectively establishes a correspondence between an underlying string and a specific surface string.

(1) a. The *candidate set* consists of triplets $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ of an *underlying* segment string $\mathbf{a}$ and a *surface* segment string $\mathbf{x}$ together with a *correspondence relation* $\rho_{\mathbf{a},\mathbf{x}}$ between the segments of $\mathbf{a}$ and those of $\mathbf{x}$.

b. A *phonological grammar* G maps a segment string $\mathbf{a}$ to a candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ whose underlying form is indeed $\mathbf{a}$.

Correspondence relations will be denoted by thin lines. To illustrate, (2) represents the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ whose underlying string $\mathbf{a}$ is /bnɪk/, whose surface string $\mathbf{x}$ is [blɪk], whose correspondence relation $\rho_{\mathbf{a},\mathbf{x}}$ maps underlying to surface segments respecting their ordering in the strings.

(2)
$$\begin{array}{rl} \mathsf{a} = & \text{b n ɪ k} \\ & \;|\;\;|\;|\;| \\ \mathsf{x} = & \text{b l ɪ k} \end{array}$$

In general, a single segment might occur multiple times in a segment string (say, the onset and the coda of a syllable might host the same consonant) and correspondence relations might need to distinguish between those occurrences. Thus, correspondence relations cannot be defined simply as relations between the two *sets* of underlying and surface segments. To keep the presentation straightforward, (1a) effectively presupposes (following common practice) that two occurrences of the same segment in a string are distinguished through diacritics (say, indices) which are visible to the correspondence relations but have no other phonological content.

Consider two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ which share a surface form x. Suppose that the underlying form b is more similar to x than the other underlying form a is. In other words, that the candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ has more *internal similarity* than the candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$. Tesar formalizes this assumption through the following condition

(3) $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq_{\mathrm{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$

where $\leq_{\mathrm{sim}}$ is a *similarity order*, namely an ordering relation among candidates (or, more precisely, among candidates which share the surface form) based on their degree of internal similarity.

Suppose that a phonological grammar G maps the less similar underlying form a to the surface form x, namely that $G(\mathsf{a}) = (\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$. This means that x is phonotactically licit and that a is not too dissimilar from x. Since the phonotactic status of x does not depend on the underlying form and furthermore b is even more similar to x, the grammar G should map also the more similar underlying form b to that same surface form x, namely $G(\mathsf{b}) = (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$. Tesar (2014, chapter 2) calls *output-driven* any phonological grammar which abides by this logic.

Definition 1. *A phonological grammar G is* output-driven *relative to a similarity order $\leq_{sim}$ provided the following implication holds*

(4) **If:** $G(\boldsymbol{\mathsf{a}}) = (\boldsymbol{\mathsf{a}}, \boldsymbol{\mathsf{x}}, \rho_{\boldsymbol{\mathsf{a}},\boldsymbol{\mathsf{x}}})$ (candidate with less internal similarity)
Then: $G(\boldsymbol{\mathsf{b}}) = (\boldsymbol{\mathsf{b}}, \boldsymbol{\mathsf{x}}, \rho_{\boldsymbol{\mathsf{b}},\boldsymbol{\mathsf{x}}})$ (candidate with more internal similarity)

for any two candidates $(\boldsymbol{\mathsf{a}}, \boldsymbol{\mathsf{x}}, \rho_{\boldsymbol{\mathsf{a}},\boldsymbol{\mathsf{x}}})$ and $(\boldsymbol{\mathsf{b}}, \boldsymbol{\mathsf{x}}, \rho_{\boldsymbol{\mathsf{b}},\boldsymbol{\mathsf{x}}})$ which share the surface form $\boldsymbol{\mathsf{x}}$ and satisfy the similarity inequality $(\boldsymbol{\mathsf{a}}, \boldsymbol{\mathsf{x}}, \rho_{\boldsymbol{\mathsf{a}},\boldsymbol{\mathsf{x}}}) \leq_{sim} (\boldsymbol{\mathsf{b}}, \boldsymbol{\mathsf{x}}, \rho_{\boldsymbol{\mathsf{b}},\boldsymbol{\mathsf{x}}})$. □

Output-drivenness is predicated on a notion of phonological similarity formalized through a similarity order $\leq_{\mathrm{sim}}$ among candidates (which share the surface form). The rest of this section focuses on the issue of the proper definition of phonological similarity.

1.2. Phonological similarity: Tesar's concrete definition

Tesar's original definition of phonological similarity is based on the notion of "*inventory of disparities*: the representational configurations that constitute individual instances of disparity" (p. 28) between an underlying string and the corresponding surface string. Tesar provisionally considers an inventory of three types of disparities, summarized in clauses (a), (b), and (c) of the following definition 2 (based on Tesar 2014, section 2.4.2). Clause (c) presupposes a phonological *feature* φ which maps segments to feature values. Throughout this paper, a feature φ can be *binary* (it takes two values) or *multi-valued* (it takes more than two values). Yet, all the features considered are required to be *total* relative to the candidate set: for every candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ in the candidate set, for every segment of the underlying string a or the surface string x, all the features considered are defined for that segment (see Magri 2018 for discussion of output-drivenness in a representational framework which allows for *partial* phonological features).

Definition 2. *Given a candidate $(\boldsymbol{\mathsf{a}}, \boldsymbol{\mathsf{x}}, \rho_{\boldsymbol{\mathsf{a}},\boldsymbol{\mathsf{x}}})$:*

(a) *An* insertion disparity *is any segment x of the surface string $\boldsymbol{\mathsf{x}}$ which has no corresponding segment in the underlying string $\boldsymbol{\mathsf{a}}$ according to the correspondence relation $\rho_{\boldsymbol{\mathsf{a}},\boldsymbol{\mathsf{x}}}$.*

(b) *A* deletion disparity *is any segment a of the underlying string $\boldsymbol{\mathsf{a}}$ which has no corresponding segment in the surface string $\boldsymbol{\mathsf{x}}$ according to the correspondence relation $\rho_{\boldsymbol{\mathsf{a}},\boldsymbol{\mathsf{x}}}$.*

(c) *A* φ-disparity *relative to some feature φ is any pair $(\mathsf{a}, \mathsf{x}) \in \rho_{\boldsymbol{\mathsf{a}},\boldsymbol{\mathsf{x}}}$ of an underlying segment a and a surface segment x corresponding through $\rho_{\boldsymbol{\mathsf{a}},\boldsymbol{\mathsf{x}}}$ but differing with respect to feature φ, namely $\varphi(\mathsf{a}) \neq \varphi(\mathsf{x})$.* □

Intuitively, a candidate counts as having more[1] internal similarity than another candidate provided every disparity relative to the former candidate comes with a *corresponding* disparity relative to the latter candidate, so that indeed the former candidate has less disparities and thus more internal similarity than the latter candidate. The following definition 3 summarizes Tesar's notion of corresponding disparities (based on Tesar 2014, section 2.4.3). For full generality, the definition is stated for two candidates which might differ also for their surface forms (although output-drivenness only compares candidates which share the same surface form). Clause (a) says that there exists a function from the segments of $\mathbf{y}$ epenthetic relative to the candidate $(\mathbf{b}, \mathbf{y}, \rho_{\mathbf{b},\mathbf{y}})$ to the segments of $\mathbf{x}$ epenthetic relative to the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ such that two different epenthetic segments of $\mathbf{y}$ cannot correspond to the same epenthetic segment of $\mathbf{x}$, whereby the function qualifies as *one-to-one*. This condition is equivalent to the condition that $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ has at least as many insertion disparities as $(\mathbf{b}, \mathbf{y}, \rho_{\mathbf{b},\mathbf{y}})$. Causes (b) and (c) are formulated analogously and ensure that $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ has at least as many deletion disparities and at least as many φ-disparities as $(\mathbf{b}, \mathbf{y}, \rho_{\mathbf{b},\mathbf{y}})$.

Definition 3. *Consider two candidates $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ and $(\boldsymbol{b}, \boldsymbol{y}, \rho_{\boldsymbol{b},\boldsymbol{y}})$ with possibly different surface forms $\boldsymbol{x}$ and $\boldsymbol{y}$:*

(a) *The insertion disparities relative to the candidate $(\boldsymbol{b}, \boldsymbol{y}, \rho_{\boldsymbol{b},\boldsymbol{y}})$ have* corresponding *insertion disparities relative to the candidate $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ provided there exists a one-to-one function from the insertion disparities of $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ to the insertion disparities of $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$.*

(b) *The deletion disparities relative to the candidate $(\boldsymbol{b}, \boldsymbol{y}, \rho_{\boldsymbol{b},\boldsymbol{y}})$ have* corresponding *deletion disparities relative to the candidate $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ provided there exists a one-to-one function from the deletion disparities of $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ to the deletion disparities of $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$.*

(c) *The φ-disparities relative to the candidate $(\boldsymbol{b}, \boldsymbol{y}, \rho_{\boldsymbol{b},\boldsymbol{y}})$ have* corresponding *φ-disparities relative to the candidate $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ provided there exists a one-to-one function from the φ-disparities of $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ to the φ-disparities of $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$.* □

Tesar submits the intuition that, "to support a claim of greater internal similarity, corresponding disparities must be *identical*: the disparities must be instances of the same disparity in the inventory of disparities" (p. 52). The following definition 4 summarizes Tesar's notion of identity between disparities (based on Tesar 2014, section 2.4.3). Following Tesar, the definition is stated for two candidates which share the same surface form. Because of the shared surface form, identity at the surface level requires segment identity: both clauses (a) and (c) require the two surface segments x' and x'' to be the same segment of the string $\mathbf{x}$. "Same" here means that x' and x'' occupy exactly the same position in the string $\mathbf{x}$. It does not suffice that they have, say, the same quality. I will come back to this issue below in subsection 2.3. Since the two candidates compared feature different underlying forms, identity at the underlying level cannot be construed as segment identity. Clause (b) for identity between deletion disparities thus uses a feature set Φ and requires the two underlying segments a and b to coincide for every feature in this set Φ while clause (c) for identity between φ-disparities requires a and b to coincide for the specific feature φ.

Definition 4. *Consider a feature set Φ and two candidates $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ and $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ sharing the surface form $\boldsymbol{x}$:*

(a) *An insertion disparity x' relative to the candidate $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ and an insertion disparity x'' relative to the candidate $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ are* identical *provided x' and x'' are the same segment of the string $\boldsymbol{x}$.*

[1] For the sake of readability, I use "more" and "less" as synonymous of "not less than" and "not more than" throughout the paper.

(b) *A deletion disparity* $\boldsymbol{a}$ *relative to the candidate* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ *and a deletion disparity* $\boldsymbol{b}$ *relative to the candidate* $(\boldsymbol{b}, \boldsymbol{y}, \rho_{\boldsymbol{b},\boldsymbol{y}})$ *are* identical *provided* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *match for every feature in the feature set* Φ.

(c) *A* φ*-disparity* (a, x') *relative to the candidate* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ *and a* φ*-disparity* (b, x'') *relative to the candidate* $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ *are* identical *provided* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *are assigned the same value by the feature* φ *and furthermore* x' *and* x'' *are the same segment of the string* $\boldsymbol{x}$. □

We are now ready to put the pieces together into the following definition 5 of phonological similarity (based on Tesar 2014, section 2.4.4).

Definition 5. *For any two candidates* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ *and* $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ *sharing the surface form* $\boldsymbol{x}$*, the former candidate is said to have* less internal similarity *than the latter candidate provided every disparity relative to the candidate* $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ *with more internal similarity admits a corresponding and identical disparity relative to the candidate* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ *with less internal similarity. In this case we write* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}}) \leq^{\Phi}_{sim} (\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$*, where* Φ *is the set of features (possibly multi-valued but all total) used to compute featural disparities (according to clause (c) of definition 2) and to determine identity between deletion disparities (according to clause (b) of definition 4).* □

The relation $\leq^{\Phi}_{\text{sim}}$ thus defined qualifies as a partial order over candidates (sharing the surface form) under the additional assumptions recalled below in subsection 3.5. I will therefore refer to $\leq^{\Phi}_{\text{sim}}$ as *Tesar's similarity order.* Subsection 1.4 provides a few examples.

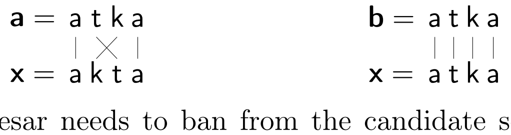
1.3. **Phonological similarity: an axiomatic definition based on faithfulness constraints**

Tesar's notion of output-drivenness is framework-independent: definition 1 above only assumes phonological grammars to be mappings from underlying forms to phonological candidates, but it does not make any specific assumptions on how phonological grammars are defined (in terms of rule ordering, constraint ranking, constraint weighting, etc.). Correspondingly, Tesar wants the definition of similarity that output-drivenness is predicated on to be framework-independent as well. Indeed, his definition 5 just recalled does not make use of any framework-specific notion and instead describes similarity *concretely*, in terms of *disparities* between strings and correspondence relations. Framework-independence is important because output-drivenness is meant to generalize the distinction between transparent and opaque phonology beyond the traditional framework-specific definition in terms of counter-feeding and counter-bleeding rule ordering.

Yet, the notion of *disparity* at the core of Tesar's definition of similarity is just a different name for a faithfulness constraint violation. Indeed, the insertion, deletion, and featural disparities formalized by Tesar's definition 2 are just a different name for the violations of the three faithfulness constraints Dep, Max, and Ident (which are indeed the only three faithfulness constraints considered by Tesar in his application of the theory of output-drivenness to OT). This is of course not surprising: faithfulness constraints are meant precisely to measure phonological similarity between underlying and corresponding surface forms. Considerations of faithfulness are intrinsic to phonological theorizing, independently of the specific framework adopted. Framework independency is therefore not an a priori impediment to a definition of phonological similarity stated in terms of faithfulness constraint violations.

Furthermore, Tesar's definition of similarity is narrowly suited to a restrictive representational framework. To illustrate, candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ in (5) has intuitively less internal similarity than the corresponding candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$, as the latter is the identity candidate which has maximal internal similarity (see below subsection 3.4) while the former displays metathesis. Yet, Tesar's notion of similarity would not be sensitive to the difference between these two candidates, as metathesis does not belong to the inventory of disparities considered by Tesar's definition 2.

(5)



Indeed, Tesar needs to ban from the candidate set any candidates with disparities other than feature mismatch, deletion, and epenthesis, because similarity is too coarsely defined (see below subsection 3.5). Consequently, the theory of output-drivenness developed by Tesar on the basis of his notion of similarity is restricted to OT typologies whose faithfulness constraint set only consists of IDENT, MAX, and DEP constraints, while other faithfulness constraints (such LINEARITY; Heinz 2005) are banned. A definition of phonological similarity stated in terms of an arbitrary set of faithfulness constraints can lead to a substantial generalization of Tesar's original theory.

Based on these considerations, Magri (2017) proposes the following alternative definition 6 of similarity in terms of the axiom (6) on the faithfulness constraints in a faithfulness constraint set $\mathcal{F}$. The notation "$\leq_{\text{sim}}^{\mathcal{F}}$" highlights the dependence of the definition on the faithfulness constraint set $\mathcal{F}$. The latter set can of course contain other faithfulness constraints besides IDENT, MAX, and DEP. If in particular $\mathcal{F}$ contains a faithfulness constraint such as LINEARITY, the corresponding notion of similarity $\leq_{\text{sim}}^{\mathcal{F}}$ is rich enough to distinguish between a pair of candidates such as those in (5), which can therefore both legitimately belong to the candidate set.

Definition 6. *Given a faithfulness constraint set $\mathcal{F}$, for any two candidates $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ and $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ sharing a surface form $\mathbf{x}$, the former candidate is said to have* less internal similarity *than the latter provided the candidate set contains a candidate $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ which puts in correspondence the two strings $\mathbf{a}$ and $\mathbf{b}$ in such a way that the inequality*[2]

$$(6) \quad F\big(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}\big) \geq F\big(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}}\big) + \underbrace{F\big(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}}\big)}_{(*)}$$

holds for every faithfulness constraint F in the faithfulness constraint set $\mathcal{F}$. In this case, we write $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq_{sim}^{\mathcal{F}} (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$. □

The relation $\leq_{\text{sim}}^{\mathcal{F}}$ thus defined qualifies as a partial order over candidates (sharing the surface form) under the additional assumptions recalled below in subsections 3.1, and 3.2, 3.3. I will therefore refer to $\leq_{\text{sim}}^{\mathcal{F}}$ as the *faithfulness-based similarity order.*

As seen above, Tesar's definition of the similarity order $\leq_{\text{sim}}^{\Phi}$ rests on the three notions of inventory of disparities, corresponding disparities, and identical disparities. The conceptual correspondence between those three notions and the alternative definition 6 of the similarity order $\leq_{\text{sim}}^{\mathcal{F}}$ can be made explicit as follows. Tesar's inventory of insertion, deletion, and φ-disparities recalled above in definition 2 can be straightforwardly translated into violations of the faithfulness constraints DEP, MAX, and IDENT$_{\varphi}$. Condition (6) entails in particular the inequality $F(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \geq F(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$. This inequality says that the less similar candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ violates the faithfulness constraints in $\mathcal{F}$ more than the more similar candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$. Since faithfulness violations correspond to Tesar's disparities, this inequality $F(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \geq F(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ captures Tesar's condition that the disparities of the more similar candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ admit corresponding disparities in the less similar candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ in the sense of definition 3 above. Finally, because of the additional term (*) (formally justified in Magri 2017, sections 5.3 and 5.4), condition (6) requires the less similar candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ to "make up" not only for any faithfulness violation incurred by the more similar candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ but also for any faithfulness violation incurred by the candidate $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ which puts in correspondence the two strings $\mathbf{a}$ and $\mathbf{b}$ being compared. This additional term (*) turns out to correspond to Tesar's additional requirement that corresponding disparities must be identical in the sense of definition 4, as clarified by the examples discussed in the next subsection.

1.4. **Examples**

This subsection illustrates the two notions of similarity $\leq_{\text{sim}}^{\Phi}$ and $\leq_{\text{sim}}^{\mathcal{F}}$ provided by definitions 5 and 6 with a few examples. The goal of these examples to is to show that, whenever a similarity inequality fails relative to Tesar's similarity $\leq_{\text{sim}}^{\Phi}$ because of the requirement that corresponding disparities be identical, it also fails relative to the new notion of similarity $\leq_{\text{sim}}^{\mathcal{F}}$ because of the additional term (*) in the inequality (6). This fact says that the additional term (*) captures Tesar's requirement that corresponding disparities be identical.

[2] As usual, a faithfulness constraint F is construed here as a function from candidates to non-negative integers and the notation "$F(candidate)$" denotes the number of violations assigned by F to the candidate considered.

To start, I notice that the two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ in (7) satisfy the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ relative to the feature set $\Phi = \{[\text{high}], [\text{voice}]\}$. In fact, the less similar candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ features two disparities: the disparity (/i/, [e]) relative to the feature $\varphi = [\text{high}]$; and the deletion disparity /d/. The candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ instead only features the deletion disparity /d/. Only the clauses (b) of definitions 3 and 4 of corresponding and identical disparities thus apply. And they are trivially satisfied by pairing up the deleted coda of b = /red/ with the identical deleted coda of a = /rid/.

(7) $$\begin{array}{llll} \mathsf{a} = & \mathsf{r} & \mathsf{i} & \mathsf{d} \\ & | & | & \\ \mathsf{x} = & \mathsf{r} & \mathsf{e} & \end{array} \qquad\qquad \begin{array}{llll} \mathsf{b} = & \mathsf{r} & \mathsf{e} & \mathsf{d} \\ & | & | & \\ \mathsf{x} = & \mathsf{r} & \mathsf{e} & \end{array}$$

These two candidates in (7) also satisfy the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ relative to the faithfulness constraint set $\mathcal{F} = \{\text{Max, Dep, Ident}_{[\text{high}]}\ \text{Ident}_{[\text{voice}]}\}$, because each of those four faithfulness constraints satisfies condition (6) when the candidate $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ which appears in (*) is chosen for instance as in (8).

(8) $$\begin{array}{llll} \mathsf{a} = & \mathsf{r} & \mathsf{i} & \mathsf{d} \\ & | & | & | \\ \mathsf{b} = & \mathsf{r} & \mathsf{e} & \mathsf{d} \end{array}$$

Let me now consider a small variant of this example, whereby the coda of the underlying form b (but not that of the underlying form a) is voiceless, as in the two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ in (9). The similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ fails in this case. In fact, both candidates feature a unique deletion disparity, namely the coda /d/ for the candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and the coda /t/ for the candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$. Any correspondence between deletion disparities (in the sense of Tesar's definition 3) must therefore hold between those two deleted codas. Yet, those two corresponding deletion disparities are not identical (in the sense of Tesar's definition 4), because the feature set Φ contains the feature [voice] and the two deleted codas differ in voicing.

(9) $$\begin{array}{llll} \mathsf{a} = & \mathsf{r} & \mathsf{i} & \mathsf{d} \\ & | & | & \\ \mathsf{x} = & \mathsf{r} & \mathsf{e} & \end{array} \qquad\qquad \begin{array}{llll} \mathsf{b} = & \mathsf{r} & \mathsf{e} & \mathsf{t} \\ & | & | & \\ \mathsf{x} = & \mathsf{r} & \mathsf{e} & \end{array}$$

These two candidates in (9) also fail at the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$. In fact, the two obvious choices for the candidate $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ which appears in (*) of the definitional condition (6) are the two candidates in (9).[3] Yet, condition (6) fails for $F = \text{Ident}_{[\text{voice}]}$ in the case of the first of those two candidates and it fails for $F = \text{Max}$ in the case of the second. Crucially, if we were to remove the term (*) from the condition (6), the resulting faithfulness inequality $F(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \geq F(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ would hold for each of the four faithfulness constraint in $\mathcal{F}$.

(10) $$\begin{array}{llll} \mathsf{a} = & \mathsf{r} & \mathsf{i} & \mathsf{d} \\ & | & | & | \\ \mathsf{b} = & \mathsf{r} & \mathsf{e} & \mathsf{t} \end{array} \qquad\qquad \begin{array}{llll} \mathsf{a} = & \mathsf{r} & \mathsf{i} & \mathsf{d} \\ & | & | & \\ \mathsf{b} = & \mathsf{r} & \mathsf{e} & \mathsf{t} \end{array}$$

In conclusion, the failure of Tesar's similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ is due to the requirement that corresponding (deletion) disparities be identical. And the failure of the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ is due to the term (*) in condition (6). This additional term (*) in condition (6) thus plays the role of the identity requirement between corresponding disparities in Tesar's original definition 5 of similarity

As another example, the two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ in (11) satisfy the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ relative to the feature set $\Phi = \{[\text{high}], [\text{place}]\}$. In fact, the less similar candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ features two disparities: the disparity (/p/, [t]) relative to the

[3] Besides the two obvious choices in (10), no other candidates $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ could help rescue the inequality (6). In fact, $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ cannot violate Dep, because the more similar candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ in (9) violates Dep and the less similar candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ violates it only once. Hence, the surface coda [t] must have an underlying correspondent in $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$. If that correspondent is the underlying coda /d/, the inequality (6) fails for $\text{Ident}_{[\text{voice}]}$. If the correspondent of the surface coda [t] is the underlying vowel /i/, the inequality (6) fails for $\text{Ident}_{[\text{syllabic}]}$. Finally, if the correspondent of the surface coda [t] is the underlying onset /r/, the inequality (6) fails for $\text{Ident}_{[\text{continuancy}]}$.

feature $\varphi =$ [place]; and the disparity (/a/, [i]) relative to the feature $\varphi =$ [high]. The candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ instead only features the disparity (/p/, [t]) relative to the feature $\varphi =$ [place]. Only the clauses (c) of definitions 3 and 4 of corresponding and identical disparities thus apply. And they are trivially satisfied by pairing up the mismatching pair (/p/, [t]) of $\rho_{\mathsf{b},\mathsf{x}}$ with the identical mismatching pair (/p/, [t]) of $\rho_{\mathsf{a},\mathsf{x}}$

```
(11)   a = p a          b = p i
           | |              | |
       x = t i          x = t i
```

These two candidates in (11) also satisfy the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ relative to the faithfulness constraint set $\mathcal{F} = \{$Max, Dep, Ident$_{\text{[high]}}$, Ident$_{\text{[place]}}\}$, because each of those four faithfulness constraints satisfies condition (6) when the candidate $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ which appears in (*) is chosen for instance as in (12).

```
(12)   a = p a
           | |
       b = p i
```

Let me now consider a small variant, whereby the onset of the underlying form b (but not that of the underlying form a) is a velar, as in the two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ in (13). The similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ fails in this case. In fact, assume that the feature [place] is multi-valued and takes three values corresponding to the three major places of articulation (see for instance de Lacy 2006, section 2.3.2.1.1). The candidates $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ and $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ feature a unique [place]-disparity, namely the pair (/p/, /t/) for the candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and the pair (/k/, /t/) for the pair $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$. Any correspondence between [place]-disparities (in the sense of Tesar's definition 3) thus needs to put those two disparities in correspondence. Yet, those two corresponding [place]-disparities are not identical (in the sense of Tesar's definition 4), because the feature [place] distinguishes between the labial and velar place of the two underlying onsets /p/ and /k/.

```
(13)   a = p a       b = k i
           | |           | |
       x = t i       x = t i
```

These two candidates in (13) also fail at the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$. In fact, the two obvious choices for the candidate $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ which appears in (*) of the definitional condition (6) are the two candidates in (14).[4] Yet, condition (6) fails for $F =$ Ident$_{\text{[place]}}$ in the case of the first candidate and it fails for $F =$ Max in the case of the second. Crucially, if we were to remove the term (*) from the condition (6), the resulting faithfulness inequality $F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) \geq F\big(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}\big)$ would hold for each of the four faithfulness constraint in $\mathcal{F}$.

```
(14)   a = p a          a = p a
           | |                |
       b = k i          b = k i
```

Once again, the failure of Tesar's similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ is due to the requirement that corresponding (featural) disparities be identical. And the failure of the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ is due to the term (*) in condition (6), further illustrating the fact that this additional term (*) in condition (6) indeed plays the role of the identity requirement between corresponding disparities in Tesar's original definition 5.

1.5. **Summary**

Subsection 1.1 has recalled Tesar's (2014) notion of output-drivenness, which formalizes the intuition that transparent (as opposed to opaque) phonological mappings "respect" phonological

[4] A reasoning analogous to that in footnote 3 shows that, besides the two obvious choices in (14), no other candidates $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ could help rescue the inequality (6).

similarity. Subsection 1.2 has reviewed Tesar's *concrete* definition 5 of similarity in terms of disparities between underlying and surface strings and their correspondence relations. Subsection 1.3 has reviewed Magri's (2017) alternative *axiomatic* definition 6 of similarity in terms of the faithfulness violation inequality (6). The examples considered in subsection 1.4 suggest that the two definitions are tightly connected. The next section makes this connection explicit.

2. Relationship between the two notions of similarity

This section shows that Tesar's concrete definition 5 of similarity falls as a special case under the axiomatic definition 6. The theory of output-drivenness reconstructed in Magri (2017) thus subsumes Tesar's (2014) original theory as a special case.

2.1. The faithfulness-based definition of similarity subsumes Tesar's definition

Tesar frames his theory of output-drivenness within a representational framework where all correspondence relations are *one-to-one*, namely they display neither breaking nor coalescence (although deletion and epenthesis are of course allowed). From now on, I thus focus on two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ whose correspondence relations $\rho_{\mathsf{a},\mathsf{x}}$ and $\rho_{\mathsf{b},\mathsf{x}}$ are both one-to-one. Suppose that the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ holds relative to the similarity order $\leq^{\Phi}_{\text{sim}}$ provided by Tesar's concrete definition 5. Proposition 1 below says that this similarity inequality entails that it is then possible to construct some one-to-one correspondence relation $\rho_{\mathsf{a},\mathsf{b}}$ between the two strings a and b such that the identity (15) holds for the core faithfulness constraints Max, Dep, and Ident. The latter identity is a special case of the inequality (6) which appears in the definition 6 of the similarity order $\leq^{\mathcal{F}}_{\text{sim}}$. Consider the triplet $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ featuring that properly constructed correspondence relation $\rho_{\mathsf{a},\mathsf{b}}$. Assume that it belongs indeed to the candidate set. That is indeed the case if, for instance, the candidate set consists of all the triplets that can be constructed out of all strings of segments drawn from a base segment set together with all one-to-one correspondence relations between any two of those strings. I can then conclude that the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ holds relative to the faithfulness constraint set $\mathcal{F}$ consisting of the three core faithfulness constraints Max, Dep, and Ident. In other words, proposition 1 establishes that the notion of similarity formalized through the relation $\leq^{\mathcal{F}}_{\text{sim}}$ subsumes the notion of similarity formalized through Tesar's relation $\leq^{\Phi}_{\text{sim}}$ when we restrict ourselves to candidates whose correspondence relations are one-to-one, as independently required by the theory of output-drivenness.

Proposition 1. *Consider two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ sharing the surface form x. Assume that both correspondence relations $\rho_{\mathsf{a},\mathsf{x}}$ and $\rho_{\mathsf{b},\mathsf{x}}$ are one-to-one. If $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{sim} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$, there exists a one-to-one correspondence relation $\rho_{\mathsf{a},\mathsf{b}}$ between the two strings a and b such that the following identity*

$$(15) \quad F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) = F\big(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}\big) + F\big(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}}\big)$$

holds when F is Max, Dep, *or* Ident$_{\varphi}$ *for any feature φ in the set Φ.* □

Perhaps surprisingly, the proof of proposition 1 is far from trivial. As discussed in the appendix, it can be derived from an intermediate result established in the very sophisticated analysis developed in Tesar (2014, chapter 3). For completeness, the appendix also provides a self-contained proof which effectively simply reboots various ingredients of Tesar's analysis. The main innovation consists of the following observation: the fact that the identity (15) holds for Max can be deduced straightforwardly from the fact that it holds for Dep, without the need to establish the identity from scratch twice for both Dep and Max. This is because the two faithfulness constraints Dep and Max are one the inverse of the other, in the sense that the following identity holds

$$(16) \quad \text{Max}\big(\mathsf{x}, \mathsf{y}, \rho_{\mathsf{x},\mathsf{y}}\big) = \text{Dep}\big(\mathsf{y}, \mathsf{x}, \rho^{-1}_{\mathsf{x},\mathsf{y}}\big)$$

for any candidate $(\mathsf{x}, \mathsf{y}, \rho_{\mathsf{x},\mathsf{y}})$ and its *inverse* candidate $(\mathsf{y}, \mathsf{x}, \rho^{-1}_{\mathsf{x},\mathsf{y}})$, namely the candidate obtained by swapping the underlying with the surface form and by putting them into correspondence through the inverse $\rho^{-1}_{\mathsf{x},\mathsf{y}}$ of the original correspondence relation $\rho_{\mathsf{x},\mathsf{y}}$.

2.2. The faithfulness-based definition of similarity generalizes Tesar's definition

Proposition 1 above ensures that Tesar's similarity order $\leq^{\Phi}_{\text{sim}}$ concretely defined in terms of disparities is subsumed under the similarity order $\leq^{\mathcal{F}}_{\text{sim}}$ defined through an axiom on the faithfulness constraints in the faithfulness constraint set $\mathcal{F}$. The two similarity orders are nonetheless different: it is easy to construct pairs of candidates $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ and $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ whose correspondence relations $\rho_{\mathbf{a},\mathbf{x}}$ and $\rho_{\mathbf{b},\mathbf{x}}$ are one-to-one such that: the similarity inequality $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ holds relative to the axiomatically defined similarity order $\leq^{\mathcal{F}}_{\text{sim}}$ corresponding to a faithfulness constraint set $\mathcal{F}$ consisting of only $\textsc{Dep}$, $\textsc{Max}$, and $\textsc{Ident}_{\varphi}$ with $\varphi \in \Phi$; yet, the similarity inequality $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq^{\Phi}_{\text{sim}} (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ fails relative to Tesar's similarity order $\leq^{\Phi}_{\text{sim}}$ corresponding to the feature set Φ. This subsection presents schematically a couple of such cases.

Consider the two candidates $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ and $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ in (17), whose correspondence relations are obviously one-to-one. Assume that corresponding segments agree for every feature in the feature set Φ (equivalently, assume that Φ is empty). The candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ then only features the insertion disparity x_1. The candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ instead features two insertion disparities x_2 and x_3. Yet, the similarity inequality $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq^{\Phi}_{\text{sim}} (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ fails. In fact, that similarity inequality would require in particular every insertion disparity of the candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ to come with a corresponding *identical* insertion disparity in the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$. Yet, definition 4 of identity between insertion disparities requires segment identity: two insertion disparities are considered identical only if they concern the same segment of the shared surface string $\mathbf{x}$. Thus, the insertion disparity x_1 of $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ admits no corresponding *identical* insertion disparity in $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$, because the segment x_1 is not epenthetic relative to $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$.

(17)
$$\begin{array}{rccc} \mathbf{a} = & \mathsf{a} & & \\ & | & & \\ \mathbf{x} = & \mathsf{x}_1 & \mathsf{x}_2 & \mathsf{x}_3 \end{array} \qquad\qquad \begin{array}{rccc} \mathbf{b} = & & \mathsf{b}_1 & \mathsf{b}_2 \\ & & | & | \\ \mathbf{x} = & \mathsf{x}_1 & \mathsf{x}_2 & \mathsf{x}_3 \end{array}$$

On the other hand, the similarity inequality $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ does hold relative to the faithfulness constraint set $\mathcal{F} = \{\textsc{Dep}, \textsc{Max}, \textsc{Ident}_{\varphi} \,|\, \varphi \in \Phi\}$, because condition (6) holds when the candidate $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ puts in correspondence the strings $\mathbf{a}$ and $\mathbf{b}$ for instance as in (18). In fact, the inequality (6) trivially holds for $\textsc{Max}$ and $\textsc{Ident}$, because they are not violated by any of the three candidates considered. Furthermore, it holds for $\textsc{Dep}$, because $\textsc{Dep}(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) = 2$ which is as large as the sum of $\textsc{Dep}(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}}) = 1$ and $\textsc{Dep}(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}}) = 1$.

(18)
$$\begin{array}{rcc} \mathbf{a} = & \mathsf{a} & \\ & | & \\ \mathbf{b} = & \mathsf{b}_1 & \mathsf{b}_2 \end{array}$$

As another example, consider the two candidates $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ and $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ in (19), whose correspondence relations are obviously one-to-one. The candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ only has a disparity (/e/, [i]) relative to the feature [high] while the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ has two such disparities. Yet, the similarity inequality $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq^{\Phi}_{\text{sim}} (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ fails. In fact, that similarity inequality would require in particular every [high]-disparity of the candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ to come with a corresponding *identical* [high]-disparity in the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$. Yet, definition 4 of identity between two featural disparities requires segment identity at the surface level: two featural disparities (a, x) and (b, x) are considered identical only if they concern the same segment x of the shared surface string $\mathbf{x}$. Thus, the [high]-disparity (/e/, [i]) of $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ admits no corresponding *identical* [high]-disparity in $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$, because it involves the third segment of the string $\mathbf{x}$ which does not enter into any featural disparity (it is rather epenthetic) relative to the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$.

(19)
$$\begin{array}{rccc} \mathbf{a} = & \mathsf{e} & \mathsf{e} & \\ & | & | & \\ \mathbf{x} = & \mathsf{i} & \mathsf{i} & \mathsf{i} \end{array} \qquad\qquad \begin{array}{rccc} \mathbf{b} = & \mathsf{i} & \mathsf{i} & \mathsf{e} \\ & | & | & | \\ \mathbf{x} = & \mathsf{i} & \mathsf{i} & \mathsf{i} \end{array}$$

On the other hand, the similarity inequality $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ does hold relative to the faithfulness constraint set $\mathcal{F} = \{\textsc{Dep}, \textsc{Max}, \textsc{Ident}_{\varphi} \,|\, \varphi \in \Phi\}$, because the inequality (6) holds when the candidate $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ puts in correspondence the strings a and b for instance as in (20). In fact, the inequality (6) trivially holds for MAX, because it is not violated by any of the three candidates considered. Furthermore, it holds for DEP, because $\text{DEP}(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) = 1$ which is as large as the sum of $\text{DEP}(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}) = 0$ and $\text{DEP}(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}}) = 1$. Finally, the inequality (6) holds for $\text{IDENT}_{[\text{high}]}$, because $\text{IDENT}_{[\text{high}]}(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) = 2$ which is as large as the sum of $\text{IDENT}_{[\text{high}]}(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}) = 1$ and $\text{IDENT}_{[\text{high}]}(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}}) = 1$.

(20)
$$\begin{array}{rccc} \mathsf{a} = & & \mathsf{e} & \mathsf{e} \\ & & | & | \\ \mathsf{b} = & \mathsf{i} & \mathsf{i} & \mathsf{e} \end{array}$$

2.3. An open problem

The crucial "culprit" in these counterexamples is that Tesar's definition 4 of identity between two disparities is tailored to two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ sharing the surface form x and takes advantage of that fact by adopting a very demanding notion of identity at the surface level, namely segment identity. For instance, consider a segment x epenthesized according to $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and a segment x' epenthesized according to $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$. Clause (a) of Tesar's definition 4 declares them *identical* insertion disparities only if x and x' are exactly the *same segment* of x. If instead they are, say, a coda and an onset, they do not count as the same insertion disparity, even if they are two segments of exactly the same quality (say, they are the same consonant). It is thus natural to entertain the following looser variant of Tesar's definition of identity between corresponding disparities, which replaces segment identity with identity of feature values. Differences with Tesar's original definition 4 are underlined. Since this alternative looser definition does not use segment identity at the surface level, it can be defined for two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{y}, \rho_{\mathsf{b},\mathsf{y}})$ with possibly different surface forms x and y.

Definition 4 (looser reformulation) *Consider a feature set Φ and two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{y}, \rho_{\mathsf{b},\mathsf{y}})$ with possibly different surface forms x and y:*

(a) *An insertion disparity x relative to the candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and an insertion disparity y relative to the candidate $(\mathsf{b}, \mathsf{y}, \rho_{\mathsf{b},\mathsf{y}})$ are* identical insertion disparities *provided x and y match for every feature in the feature set Φ.*

(b) *A deletion disparity a relative to the candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and a deletion disparity b relative to the candidate $(\mathsf{b}, \mathsf{y}, \rho_{\mathsf{b},\mathsf{y}})$ are* identical deletion disparities *provided a and b match for every feature in the feature set Φ.*

(c) *For any feature φ in Φ, a φ-disparity (a, x) relative to the candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and a φ-disparity (b, y) relative to the candidate $(\mathsf{b}, \mathsf{y}, \rho_{\mathsf{b},\mathsf{y}})$ are* identical φ-disparities *provided a and b are assigned the same value by feature φ and furthermore x and y match for every feature in the feature set Φ.* □

Let me denote by $\leq_{\text{sim}}^{\text{loose},\Phi}$ the similarity order defined in terms of identical corresponding disparities exactly as in Tesar's definition 5, only using the looser definition of identity above, rather than Tesar's original definition 4. The preceding considerations suggest that $\leq_{\text{sim}}^{\text{loose},\Phi}$ (just as Tesar's original $\leq_{\text{sim}}^{\Phi}$) should qualify as a special case of $\leq_{\text{sim}}^{\mathcal{F}}$, yielding a slight generalization of proposition 1. Yet, the proof of proposition 1 presented in the appendix does not extend in any obvious way from $\leq_{\text{sim}}^{\Phi}$ to $\leq_{\text{sim}}^{\text{loose},\Phi}$. I thus have to leave as an open problem the proper characterization of the relationship between the two similarity orders $\leq_{\text{sim}}^{\text{loose},\Phi}$ and $\leq_{\text{sim}}^{\mathcal{F}}$.

2.4. Summary

Tesar defines phonological similarity concretely in terms of disparities between segment strings and correspondence relations. His inventory of disparities only contains insertion, deletion and featural disparities. As a consequence, Tesar's application of his theory of output-drivenness to OT is limited to constraint sets which only contain DEP, MAX, and IDENT faithfulness constraints. How can the theory be extended beyond this restrictive theory of faithfulness? How can a larger inventory of disparities be taken into account? A natural strategy is to interpret *disparities* as faithfulness constraint violations and thus to axiomatize Tesar's notion of similarity through a condition on the faithfulness constraint violations. This section has shown that the faithfulness inequality (6) used in definition 6 indeed provides one such faithfulness axiomatization of Tesar's notion of similarity.

3. Formal properties of the two notions of similarity

Output-drivenness requires a way to compare candidates in terms on their degree of internal similarity. Intuitively, any such comparison should yield a partial *ordering* of the candidates (sharing the surface form) based on their internal similarity. As motivated in the preceding section, I propose to formalize this comparison as the relation $\leq^{\mathcal{F}}_{\text{sim}}$ provided by definition 6. In order for this proposal to make sense, this relation $\leq^{\mathcal{F}}_{\text{sim}}$ thus needs to qualify as a partial order over the candidate set. Subsections 3.1-3.3 thus probe deeper into the formal properties of the relation $\leq^{\mathcal{F}}_{\text{sim}}$ with the goal of establishing conditions on the candidate set and the faithfulness constraint set $\mathcal{F}$ which suffice to ensure that $\leq^{\mathcal{F}}_{\text{sim}}$ is indeed a partial order. Subsection 3.5 then compares the conditions thus obtained with the conditions needed to ensure that the relation $\leq^{\Phi}_{\text{sim}}$ provided by Tesar's definition 5 qualifies as a partial order.

3.1. **Reflexivity**

Let me investigate whether the relation $\leq^{\mathcal{F}}_{\text{sim}}$ provided by definition 6 is *reflexive*, namely whether it satisfies condition (21) for every candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$.

(21) $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$

To this end, let me qualify the representational framework through the *U-reflexivity* axiom (22): each underlying string a is required to figure in the candidate which puts it in correspondence with itself (construed as a surface form) through the *identity correspondence relation* $\mathbb{I}_{\mathsf{a},\mathsf{a}}$ which puts each segment in correspondence with itself. This axiom requires in particular each underlying form to also count as a surface form, thus partially blurring the distinction between the two representational levels (for discussion, see Moreton 2004).

(22) If the candidate set contains a candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ whose underlying string is a, then it also contains the *identity* candidate $(\mathsf{a}, \mathsf{a}, \mathbb{I}_{\mathsf{a},\mathsf{a}})$, where $\mathbb{I}_{\mathsf{a},\mathsf{a}}$ is the identity correspondence relation among the segments of a.

The identity candidates postulated in (22) allow for a straightforward axiomatization of the notion of faithfulness constraint: a constraint F counts as a faithfulness constraint provided it assigns no violations to any identity candidates, as stated in (23). Of course this definition is substantial only because of the existence of identity candidates guaranteed by (22).

(23) $F(\mathsf{a}, \mathsf{a}, \mathbb{I}_{\mathsf{a},\mathsf{a}}) = 0$

The reflexivity condition (21) now easily follows. In fact, consider the identity candidate $(\mathsf{a}, \mathsf{a}, \mathbb{I}_{\mathsf{a},\mathsf{a}})$, whose existence is guaranteed by the U-reflexivity axiom (22). The identity (24) holds for any faithfulness constraint F, as F assign zero violations to the identity candidate $(\mathsf{a}, \mathsf{a}, \mathbb{I}_{\mathsf{a},\mathsf{a}})$ by (23).

(24) $F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) = F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}\big) + F\big(\mathsf{a}, \mathsf{a}, \mathbb{I}_{\mathsf{a},\mathsf{a}}\big)$

This identity (24) entails the inequality (6) in the definition 6 of the relation $\leq^{\mathcal{F}}_{\text{sim}}$ in the special case $\mathsf{a} = \mathsf{b}$, thus ensuring the validity of the reflexivity condition (21).

3.2. **Transitivity**

Let me investigate whether the relation $\leq^{\mathcal{F}}_{\text{sim}}$ is *transitive*, namely whether it satisfies the implication (25) for any three candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$, $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$, and $(\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}})$: if the first candidate has less internal similarity than the second which in turn has less internal similarity than the third, then the first candidate ought to have less internal similarity than the third.

(25) **If:** $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}})$
Then: $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}})$

Definition 6 establishes the relation $\leq^{\mathcal{F}}_{\text{sim}}$ between two candidates in terms of the faithfulness constraint violation inequality (6). This inequality crucially involves an additional candidate which puts into correspondence the two strings which play the role of the underlying forms in the two candidates being compared. Thus, the two similarity inequalities in the antecedent of the transitivity implication (25) guarantee in particular the existence of two correspondence relations $\rho_{\mathbf{a},\mathbf{b}}$ and $\rho_{\mathbf{b},\mathbf{c}}$ between the two pairs of strings $\mathbf{a}, \mathbf{b}$ and $\mathbf{b}, \mathbf{c}$. And the consequent of the transitivity implication (25) requires in particular the existence of a correspondence relation $\rho_{\mathbf{a},\mathbf{c}}$ between the two strings $\mathbf{a}, \mathbf{c}$. In order for the transitivity implication to hold, the existence of the two correspondence relations $\rho_{\mathbf{a},\mathbf{b}}$ and $\rho_{\mathbf{b},\mathbf{c}}$ must therefore entail the existence of a correspondence relation $\rho_{\mathbf{a},\mathbf{c}}$. Here is a way to implement this idea.

Consider two candidates $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ and $(\mathbf{b}, \mathbf{c}, \rho_{\mathbf{b},\mathbf{c}})$ which share a string $\mathbf{b}$ as the surface and underlying form respectively, as represented in (26). The *composition* $\rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}}$ of the correspondence relations $\rho_{\mathbf{a},\mathbf{b}}$ and $\rho_{\mathbf{b},\mathbf{c}}$ is defined as follows: two segments a and c of the strings $\mathbf{a}$ and $\mathbf{c}$ are in correspondence through $\rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}}$ if and only if there exists a "mediating" segment b of the string $\mathbf{b}$ such that a corresponds to b through $\rho_{\mathbf{a},\mathbf{b}}$ and b corresponds to c through $\rho_{\mathbf{b},\mathbf{c}}$.

(26)
$$\begin{tikzcd} \mathbf{a} \arrow[r, "\rho_{\mathbf{a},\mathbf{b}}"] \arrow[rr, bend right, "\rho_{\mathbf{a},\mathbf{c}} = \rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}}"'] & \mathbf{b} \arrow[r, "\rho_{\mathbf{b},\mathbf{c}}"] & \mathbf{c} \end{tikzcd}$$

As a further qualification of the representational framework (1), I assume that the candidate set satisfies the *transitivity axiom* (27) which ensures the existence of *composition* candidates.

(27) If the candidate set contains two candidates $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ and $(\mathbf{b}, \mathbf{c}, \rho_{\mathbf{b},\mathbf{c}})$, it also contains their *composition candidate* $(\mathbf{a}, \mathbf{c}, \rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}})$.

This assumption (27) formalizes the following intuition. A surface form counts as a candidate of an underlying form provided the former can be obtained from the latter through a series of sensible phonological operations, such as deletion, epenthesis, metathesis, or modification of some feature values. If the surface form $\mathbf{b}$ can be obtained from the underlying form $\mathbf{a}$ through a series of such operations and furthermore if the surface form $\mathbf{c}$ can be obtained from $\mathbf{b}$ (construed as an underlying form) through some other series of such operations, one would then expect to be able to derive $\mathbf{c}$ directly from $\mathbf{a}$ by chaining the two series of operations.

Because of the transitivity axiom (27), the existence of the two correspondence relations $\rho_{\mathbf{a},\mathbf{b}}$ and $\rho_{\mathbf{b},\mathbf{c}}$ ensures the existence of the composition correspondence relation $\rho_{\mathbf{a},\mathbf{c}} = \rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}}$. In order to secure the transitivity implication (25), we also need the faithfulness violations assigned to the composition candidate $(\mathbf{a}, \mathbf{c}, \rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}})$ to be suitably related to the violations assigned to the two candidates $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$, $(\mathbf{b}, \mathbf{c}, \rho_{\mathbf{b},\mathbf{c}})$. Here is a way to achieve that. Faithfulness intuitively measures the "phonological distance" between underlying and corresponding surface forms. A *distance* maps two points A and B to a non-negative value $dist(A, B)$. In order to capture the intuitive notion of distance, this mapping needs to satisfy some core axioms (Rudin 1953, ch. 2). One of these axioms is the *triangle inequality* (28): the distance between two points A and C is never larger than the sum of the distance between A and B plus the distance between B and C, no matter how the intermediate point B is chosen. In other words, no side $\overline{AC}$ of a triangle can be longer than the sum $\overline{AB} + \overline{BC}$ of the other two sides.

(28) $dist(A, C) \leq dist(A, B) + dist(B, C)$

Motivated by the intuition that a faithfulness constraint F measures the phonological distance between underlying and surface forms from a certain specific perspective, Magri (2017) says that F satisfies the *faithfulness triangle inequality* (FTI) provided condition (29) holds for any two candidates $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ and $(\mathbf{b}, \mathbf{c}, \rho_{\mathbf{b},\mathbf{c}})$ and their composition candidate $(\mathbf{a}, \mathbf{c}, \rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}})$, whose existence is guaranteed by the transitivity axiom (27).

(29) $F\big(\mathbf{a}, \mathbf{c}, \rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}}\big) \leq F\big(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}}\big) + F\big(\mathbf{b}, \mathbf{c}, \rho_{\mathbf{b},\mathbf{c}}\big)$

Assume that the candidate set satisfies the transitivity axiom (27) and only contains one-to-one correspondence relations. Magri (2017, proposition 4) shows that the vast majority of the faithfulness constraints used in the phonological literature (including in particular segmental Max and Dep; featural $\textsc{Max}_{[\pm\varphi]}$ and $\textsc{Dep}_{[\pm\varphi]}$; $\textsc{Ident}_{\varphi}$ corresponding to a total feature φ; the local disjunction of any two identity constraints; Linearity, MaxLinearity, and DepLinearity; I/O-Adjacency, etcetera) satisfy the FTI.[5] Furthermore, Magri (2017, proposition 11) establishes an equivalence result between the FTI and the condition on the faithfulness constraints established by Tesar as sufficient for output-drivenness in OT. This equivalence holds when output-drivenness is construed relative to $\leq^{\mathcal{F}}_{\text{sim}}$ and the faithfulness constraints in the faithfulness constraint set $\mathcal{F}$ all comply with (a slightly stronger version of) McCarthy's (2003) *categoricity conjecture.*

Based on these considerations, let me assume that each faithfulness constraint in the faithfulness constraint set $\mathcal{F}$ satisfies the FTI and let me show that the corresponding relation $\leq^{\mathcal{F}}_{\text{sim}}$ then satisfies the transitivity implication (25). By definition 6 of the relation $\leq^{\mathcal{F}}_{\text{sim}}$, the antecedent of the transitivity implication means that there exist two candidates $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ and $(\mathbf{b}, \mathbf{c}, \rho_{\mathbf{b},\mathbf{c}})$ which validate the two inequalities (30) for every faithfulness constraint F in the set $\mathcal{F}$.

(30) a. $F\big(\mathbf{a}, \mathsf{x}, \rho_{\mathbf{a},\mathbf{x}}\big) \geq F\big(\mathbf{b}, \mathsf{x}, \rho_{\mathbf{b},\mathbf{x}}\big) + F\big(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}}\big)$

b. $F\big(\mathbf{b}, \mathsf{x}, \rho_{\mathbf{b},\mathbf{x}}\big) \geq F\big(\mathsf{c}, \mathsf{x}, \rho_{\mathbf{c},\mathbf{x}}\big) + F\big(\mathbf{b}, \mathsf{c}, \rho_{\mathbf{b},\mathbf{c}}\big)$

Consider the candidate $(\mathbf{a}, \mathsf{c}, \rho_{\mathbf{a},\mathbf{b}} \circ \rho_{\mathbf{b},\mathbf{c}})$ which is the composition of the two candidates $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ and $(\mathbf{b}, \mathsf{c}, \rho_{\mathbf{b},\mathbf{c}})$, whose existence is guaranteed by the transitivity axiom (27). The chain of inequalities (31) then holds. In step (31a), I have used (30a). In step (31b), I have used (30b). Finally in step (31c), I have used the FTI (29).

(31)
$$\begin{aligned} F\big(\mathbf{a}, \mathsf{x}, \rho_{\mathbf{a},\mathbf{x}}\big) &\overset{(a)}{\geq} F\big(\mathbf{b}, \mathsf{x}, \rho_{\mathbf{b},\mathbf{x}}\big) + F\big(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}}\big) \\ &\overset{(b)}{\geq} F\big(\mathsf{c}, \mathsf{x}, \rho_{\mathbf{c},\mathbf{x}}\big) + F\big(\mathbf{b}, \mathsf{c}, \rho_{\mathbf{b},\mathbf{c}}\big) + F\big(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}}\big) \\ &\overset{(c)}{\geq} F\big(\mathsf{c}, \mathsf{x}, \rho_{\mathbf{c},\mathbf{x}}\big) + F\big(\mathbf{a}, \mathsf{c}, \rho_{\mathbf{b},\mathbf{c}} \circ \rho_{\mathbf{b},\mathbf{c}}\big) \end{aligned}$$

The inequality obtained in (31) ensures that $(\mathbf{a}, \mathsf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{c}, \mathsf{x}, \rho_{\mathbf{c},\mathbf{x}})$, as required by the consequent of the transitivity implication (25).

The assumption that the faithfulness constraints in $\mathcal{F}$ all satisfy the triangle inequality is crucial in order for the relation $\leq^{\mathcal{F}}_{\text{sim}}$ to satisfy the transitivity axiom (25). Here is a counterexample. Consider the constraint F defined in (32) as the squared version of the familiar constraint Dep. Obviously, F qualifies as a *faithfulness* constraint, namely it assigns zero violations to the identity candidates and thus complies with condition (23).

(32) $$F\big(\mathbf{a}, \mathsf{x}, \rho_{\mathbf{a},\mathbf{x}}\big) = \Big(\textsc{Dep}\big(\mathbf{a}, \mathsf{x}, \rho_{\mathbf{a},\mathbf{x}}\big)\Big)^2$$

Because of the quadratic dependence on the number of epenthetic segments, the constraint F is easily seen to fail at the FTI. Suppose that the faithfulness constraint set $\mathcal{F}$ consists of only this faithfulness constraint F. The corresponding relation $\leq^{\mathcal{F}}_{\text{sim}}$ fails at the transitivity implication (25), as shown for instance by the three candidates $(\mathbf{a}, \mathsf{x}, \rho_{\mathbf{a},\mathbf{x}})$, $(\mathbf{b}, \mathsf{x}, \rho_{\mathbf{b},\mathbf{x}})$, and $(\mathsf{c}, \mathsf{x}, \rho_{\mathbf{c},\mathbf{x}})$ in (33).

(33)
$$\begin{array}{rllll} \mathbf{a} = & \mathsf{a} & & & \\ & | & & & \\ \mathbf{x} = & \mathsf{x} & \mathsf{x} & \mathsf{x} & \mathsf{x} \end{array} \qquad \begin{array}{rlllll} \mathbf{b} = & \mathsf{b} & \mathsf{b} & \mathsf{b} & \\ & | & | & & \\ \mathbf{x} = & \mathsf{x} & \mathsf{x} & \mathsf{x} & \mathsf{x} \end{array} \qquad \begin{array}{rllllll} \mathbf{c} = & \mathsf{c} & \mathsf{c} & \mathsf{c} & \mathsf{c} & \mathsf{c} \\ & | & | & | & | & \\ \mathbf{x} = & \mathsf{x} & \mathsf{x} & \mathsf{x} & \mathsf{x} & \end{array}$$

In fact, the similarity inequality $(\mathbf{a}, \mathsf{x}, \rho_{\mathbf{a},\mathbf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathbf{b}, \mathsf{x}, \rho_{\mathbf{b},\mathbf{x}})$ holds, because the constraint violation inequality (34) holds relative to the specific candidate $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ considered.

(34) $$\underbrace{F\big(\mathbf{a}, \mathsf{x}, \rho_{\mathbf{a},\mathbf{x}}\big)}_{=9} \geq \underbrace{F\big(\mathbf{b}, \mathsf{x}, \rho_{\mathbf{b},\mathbf{x}}\big)}_{=4} + \underbrace{F\big(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}}\big)}_{=4} \qquad \text{where} \qquad \begin{array}{rlll} \mathbf{a} = & \mathsf{a} & & \\ & | & & \\ \mathbf{b} = & \mathsf{b} & \mathsf{b} & \mathsf{b} \end{array}$$

[5] On the contrary, Contiguity and the local conjunction of two conjoinable faithfulness constraints fail at the FTI.

Furthermore, the similarity inequality $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}})$ also holds, because the constraint violation inequality (35) holds relative to the specific candidate $(\mathsf{b}, \mathsf{c}, \rho_{\mathsf{b},\mathsf{c}})$ considered.

$$(35) \quad \underbrace{F\big(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}\big)}_{=4} \geq \underbrace{F\big(\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}}\big)}_{=0} + \underbrace{F\big(\mathsf{b}, \mathsf{c}, \rho_{\mathsf{b},\mathsf{c}}\big)}_{=4} \qquad \text{where} \quad \begin{array}{rlllll} \mathsf{b} = & \mathsf{b} & \mathsf{b} & \mathsf{b} & & \\ & | & | & | & & \\ \mathsf{c} = & \mathsf{c} & \mathsf{c} & \mathsf{c} & \mathsf{c} & \mathsf{c} \end{array}$$

Since $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}})$, the transitivity implication (25) requires that also $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}})$. Yet, the latter similarity inequality fails. To see that, let's start by noting that the constraint violation inequality (36) fails relative to the specific candidate $(\mathsf{a}, \mathsf{c}, \rho_{\mathsf{a},\mathsf{c}})$ considered.

$$(36) \quad \underbrace{F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big)}_{=9} \not\geq \underbrace{F\big(\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}}\big)}_{=0} + \underbrace{F\big(\mathsf{a}, \mathsf{c}, \rho_{\mathsf{a},\mathsf{c}}\big)}_{=16} \qquad \text{where} \quad \begin{array}{rlllll} \mathsf{a} = & \mathsf{a} & & & & \\ & | & & & & \\ \mathsf{c} = & \mathsf{c} & \mathsf{c} & \mathsf{c} & \mathsf{c} & \mathsf{c} \end{array}$$

This of course does not suffice to conclude that the similarity inequality $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{c}, \mathsf{x}, \rho_{\mathsf{c},\mathsf{x}})$ fails, because the faithfulness violation inequality (36) could hold for a different, more careful choice of the candidate $(\mathsf{a}, \mathsf{c}, \rho_{\mathsf{a},\mathsf{c}})$. Yet, recall from subsection 2.1 that the theory of output-drivenness in OT requires the candidate set to display no breaking nor coalescence. This means that we are allowed to consider only two variants of the specific candidate $(\mathsf{a}, \mathsf{c}, \rho_{\mathsf{a},\mathsf{c}})$ in (36). First, we can consider the variant where the unique segment of the string a is in correspondence with a segment of the string c different from the initial segment. But this modification will not change the number of violations and will therefore not rescue the inequality (36). Second, we can drop the unique correspondence line and thus assume that the unique segment of the string a is in correspondence with no segment of the string c. But this modification will only increase the number of violations assigned to the candidate $(\mathsf{a}, \mathsf{c}, \rho_{\mathsf{a},\mathsf{c}})$ and will therefore not help with the inequality (36).

3.3. Antisymmetry

Let me now investigate whether the relation $\leq^{\mathcal{F}}_{\text{sim}}$ provided by definition 6 is *antisymmetric*, namely whether it satisfies the implication (37) for any two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$: if each of them has at least as much internal similarity as the other, they are the same candidate.

(37) **If:** $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$
Then: $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) = (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$.

To appreciate the problem posed by antisymmetry, suppose that $\mathcal{F}$ is empty: it contains no faithfulness constraints at all. In this extreme case, the antecedent of the antisymmetry implication (37) holds for any two arbitrary candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ and the implication thus fails. In order to secure the antisymmetry implication (37), I thus need to make sure that the faithfulness constraint set $\mathcal{F}$ used to compute similarity through $\leq^{\mathcal{F}}_{\text{sim}}$ is sufficiently rich to ensure a sufficiently strong antecedent. Thus, let me assume that the faithfulness constraint set $\mathcal{F}$ is *complete* (relative to the candidate set): for any two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ which share the surface form and yet are different because they have different underlying forms and/or different correspondence relations, the faithfulness constraint set $\mathcal{F}$ contains at least one faithfulness constraint F which acknowledges the difference by assigning different numbers of violations to those two candidates. Completeness is a rather mild assumption, as it simply formalizes the reasonable requirement that the faithfulness constraint set needs to be "commensurate" with the candidate set.[6]

Let me show that the similarity relation $\leq^{\mathcal{F}}_{\text{sim}}$ indeed satisfies the antisymmetry implication (37) when the faithfulness constraint set $\mathcal{F}$ is complete. By the definition 6 of $\leq^{\mathcal{F}}_{\text{sim}}$, the antecedent of

[6] Completeness does not contradict the condition (23) that faithfulness constraints assign the same number of violations (namely 0) to all identity candidates. In fact, completeness only looks at different candidates which share the surface form. If two identity candidates share the surface form, they are the same candidate.

the antisymmetry implication (37) means that there exist two candidates $(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}})$ and $(\mathsf{b}, \mathsf{a}, \rho_{\mathsf{b},\mathsf{a}})$ which validate the inequalities (38) for every faithfulness constraint F in the set $\mathcal{F}$.

(38) a. $F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) \geq F\big(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}\big) + F\big(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}}\big)$
b. $F\big(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}\big) \geq F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) + F\big(\mathsf{b}, \mathsf{a}, \rho_{\mathsf{b},\mathsf{a}}\big)$

The chain of inequalities in (39) then holds. In step (39a), I have used (38a) together with the fact that constraint violations are non-negative. In step (39b), I have used (38b).

(39) $F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) \overset{(a)}{\geq} F\big(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}\big) \overset{(b)}{\geq} F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) + F\big(\mathsf{b}, \mathsf{a}, \rho_{\mathsf{b},\mathsf{a}}\big)$

The inequality (39) thus derived, together with the non-negativity of constraint violations, implies that $F\big(\mathsf{b}, \mathsf{a}, \rho_{\mathsf{b},\mathsf{a}}\big) = 0$. An analogous reasoning shows that $F\big(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}}\big) = 0$. The two inequalities (38) thus become $F(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \geq F(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ and $F(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}) \geq F(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$, whereby (40).

(40) $F(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) = F(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$

Since the identity (40) holds for every faithfulness constraint F in $\mathcal{F}$ and since the faithfulness constraint set $\mathcal{F}$ is complete relative to the candidate set, I conclude that the candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ are identical, as required by the consequent of the antisymmetry implication (37).

3.4. Identity candidates have maximal internal similarity

The U-reflexivity axiom (22) requires the candidate set to contain the identity candidate corresponding to any underlying form. This requirement is complemented by the following S-reflexivity axiom (41), which requires the candidate set to also contain the identity candidate corresponding to any surface form. This axiom requires in particular each surface form to also count as an underlying form, thus further blurring the distinction between the two representational levels.

(41) If the candidate set contains a candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ whose surface string is x, then it also contains the *identity* candidate $(\mathsf{x}, \mathsf{x}, \mathbb{I}_{\mathsf{x},\mathsf{x}})$, where $\mathbb{I}_{\mathsf{x},\mathsf{x}}$ is the identity correspondence relation among the segments of x.

Intuitively, any string x is more similar to itself than to any other string a. In other words, identity candidates intuitively have the greatest internal similarity. In order for the relation $\leq^{\mathcal{F}}_{\text{sim}}$ to properly formalize the notion of phonological similarity, it thus needs to satisfy condition (42) for any candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and the corresponding identity candidate $(\mathsf{x}, \mathsf{x}, \mathbb{I}_{\mathsf{x},\mathsf{x}})$, whose existence is guaranteed by the S-reflexivity axiom (41). Whenever the similarity inequality (42) holds for every candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$, output-drivenness guarantees that phonotactically licit forms (namely forms which are the surface realization of some underlying form) are faithfully realized (when construed as underlying forms).

(42) $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\mathcal{F}}_{\text{sim}} (\mathsf{x}, \mathsf{x}, \mathbb{I}_{\mathsf{x},\mathsf{x}})$

Indeed, this similarity inequality (42) holds. In fact, every faithfulness constraint F assigns zero violations to the identity candidate $(\mathsf{x}, \mathsf{x}, \mathbb{I}_{\mathsf{x},\mathsf{x}})$, by (23). The identity (43) thus trivially holds for every faithfulness constraint F.

(43) $F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) = F\big(\mathsf{x}, \mathsf{x}, \mathbb{I}_{\mathsf{x},\mathsf{x}}\big) + F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big)$

This identity (43) entails the inequality (6) in definition 6 of the relation $\leq^{\mathcal{F}}_{\text{sim}}$ in the special case $\mathsf{b}{=}\mathsf{x}$, thus ensuring the validity of the desired condition (42).

3.5. Comparison with Tesar's notion of similarity

It is trivial to verify that the relation $\leq^{\Phi}_{\text{sim}}$ provided by Tesar's definition 5 is reflexive and transitive and that it furthermore satisfies the intuitive condition (42) that identity candidates have maximal internal similarity. The case of antisymmetry requires more care. Tesar (subsection 2.4.2) shows indeed that the antisymmetry of $\leq^{\Phi}_{\text{sim}}$ requires three additional conditions (besides the assumption that all correspondence relations in the candidate set are one-to-one). The first additional condition is that the feature set Φ is *complete* in the sense that:

(44) First additional condition:
for any two different segments (which occur in any string in any candidate), at least one feature in Φ assigns them a different value.

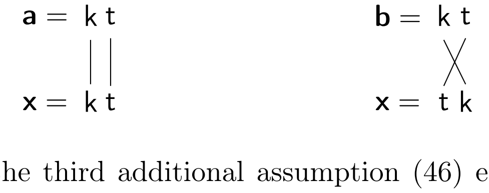
The second additional condition is that the correspondence relation $\rho_{\mathsf{a},\mathsf{x}}$ of any candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ respects the linear order of the strings a and x, in the sense that:

(45) Second additional condition:
if $(\mathsf{a}_1, \mathsf{x}_1), (\mathsf{a}_2, \mathsf{x}_2) \in \rho_{\mathsf{a},\mathsf{x}}$ and a_1 precedes a_2 in a, then x_1 precedes x_2 in x.

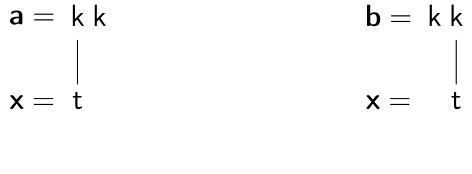
The third additional condition concerns the notion of correspondence between disparities. The assumption $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ means in particular that each deletion disparity of the candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ with more internal similarity admits a corresponding deletion disparity in the candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ with less internal similarity. By clause (b) of definition 3, this means that there exists a one-to-one function $i_{\mathsf{b},\mathsf{a}}$ from the segments of b deleted relative to $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ to the segments of a deleted relative to $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$. The composition $\left(\rho_{\mathsf{b},\mathsf{x}} \circ \rho^{-1}_{\mathsf{a},\mathsf{x}}\right)$ of $\rho_{\mathsf{b},\mathsf{x}}$ with the inverse $\rho^{-1}_{\mathsf{a},\mathsf{x}}$ of $\rho_{\mathsf{a},\mathsf{x}}$ is another relation between the segments of b and those of a. Antisymmetry of $\leq^{\Phi}_{\text{sim}}$ requires the clause (b) of the definition 3 of corresponding deletion disparities to be strengthened through the requirement that the union of the two relations $i_{\mathsf{b},\mathsf{a}}$ and $\left(\rho_{\mathsf{b},\mathsf{x}} \circ \rho^{-1}_{\mathsf{a},\mathsf{x}}\right)$ respect the linear order of the strings a and b, in the sense that:

(46) Third additional condition:
if $(\mathsf{b}_1, \mathsf{a}_1), (\mathsf{b}_2, \mathsf{a}_2) \in i_{\mathsf{b},\mathsf{a}} \cup \left(\rho_{\mathsf{b},\mathsf{x}} \circ \rho^{-1}_{\mathsf{a},\mathsf{x}}\right)$ and b_1 precedes b_2 in b, then a_1 precedes a_2 in a.

The fact that these three conditions are sufficient to ensure the antisymmetry of Tesar's similarity relation $\leq^{\Phi}_{\text{sim}}$ is not surprising in light of the results obtained so far. In fact, suppose that the antecedent of the antisymmetry implication (37) holds relative to $\leq^{\Phi}_{\text{sim}}$, namely that $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$. As seen in section 2, these similarity inequalities then hold also with $\leq^{\Phi}_{\text{sim}}$ replaced by $\leq^{\mathcal{F}}_{\text{sim}}$ when the faithfulness constraint set $\mathcal{F}$ is defined by $\mathcal{F} = \{\textsc{Dep}, \textsc{Max}, \textsc{Ident}_{\varphi} \,|\, \varphi \in \Phi\}$. As seen in subsection 3.3, to conclude that $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ are the same candidate as required by antisymmetry, we then only need to ensure that this specific faithfulness constraint set $\mathcal{F}$ is complete: if the two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ were different, then at least one of the faithfulness constraints in $\mathcal{F}$ would assign them a different number of violations. The three additional assumptions (44), (45), and (46) ensure precisely that. In fact, the first additional assumption (44) assures that $\mathcal{F}$ contains enough featural identity faithfulness constraints to pull apart any relevant difference in segmental quality. Furthermore, the second additional assumption (45) ensures that we do not need to worry about two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ as in (47): although the specific faithfulness constraint set $\mathcal{F}$ considered fails to distinguish between these two candidates, that failure is harmless because the candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ cannot belong to the candidate set because its correspondence relation does not respect the linear order of the corresponding segments.

(47)


Finally, the third additional assumption (46) ensures that we do not need to worry about two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ with $\mathsf{a} = \mathsf{b}$ as in (48): although the specific faithfulness constraint set $\mathcal{F}$ considered fails to distinguish between these two candidates, that failure is harmless because these two candidates $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ and $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ are not comparable relative to $\leq^{\Phi}_{\text{sim}}$ because the deleted segment is ordered differently with respect to the non-deleted segment in the two candidates, whereby condition (46) is easily shown to fail.

(48)


3.6. **Summary**

The results obtained in this section can be summarized into the following proposition, which provides conditions on the candidate set and on the faithfulness constraint set $\mathcal{F}$ in order for the relation $\leq^{\mathcal{F}}_{\text{sim}}$ to qualify as a partial order, as required by the intuitive notion of phonological similarity that $\leq^{\mathcal{F}}_{\text{sim}}$ is meant to capture.

Proposition 2. *Assume the candidate set (1) satisfies the reflexivity axioms (22) and (41) as well as the transitivity axiom (27). Consider a faithfulness constraint set $\mathcal{F}$ such that:*

(a) *$\mathcal{F}$ is complete relative to the candidate set: for any two different candidates $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$, $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ sharing the surface form, there exists a faithfulness constraint F in $\mathcal{F}$ which assigns them a different number of violations.*

(b) *$\mathcal{F}$ consists of faithfulness constraints which all satisfy the FTI (29).*

The corresponding relation $\leq^{\mathcal{F}}_{sim}$ provided by definition 6 is reflexive, antisymmetric, and transitive, namely it is a partial order over the candidate set; furthermore it satisfies the intuitive condition (42) that identity candidates have maximal internal similarity. □

As commented above, the completeness assumption (a) formalizes commensurabilty between the candidate set and the faithfulness constraint set. And the FTI assumption (b) formalizes the intuition that faithfulness constraints measure phonological distance in a sensible way, namely in compliance with the triangle inequality. Furthermore, the FTI is equivalent to Tesar's sufficient faithfulness condition for OT output-drivenness (under McCarthy's 2003 categoricity conjecture).

4. Conclusions

The distinction between *transparent* and *opaque* phonology is traditionally drawn in terms of counter-feeding and counter-bleeding rule ordering. How can that distinction be extended to a phonological framework which does not rely on ordered rules? In his seminal work, Tesar proposes that the distinction between transparent and opaque phonology can be equated with the framework-independent distinction between output-driven and non-output-driven phonology. The intuition is that only transparent/output-driven phonology respects phonological similarity. The development of the theory of output-drivenness thus requires an explicit notion of phonological similarity. Tesar (2014) defines similarity concretely in terms of segment strings and correspondence relations. Because of this concreteness, the resulting notion of output-drivenness can unfortunately only be applied in OT to cases where the faithfulness constraint set is very restricted, namely it only consists of Dep, Max, and Ident. To overcom this limitation, Magri (2017) defines phonological similarity axiomatically in terms of an inequality on the number of violations assigned by a given but arbitrary set of faithfulness constraints. This paper has probed deeper into the formal properties of the latter faithfulness-based definition of similarity. In particular, it has established that the faithfulness-based definition generalizes Tesar's original concrete definition of similarity: the latter indeed represents a special case of the former. Furthermore, it has shown that also the faithfulness-based definition, just as Tesar's original definition, yields a reflexive, antisymmetric and transitive relation over the candidate set, namely a partial ordering of the candidates based on their degree of internal similarity. The theory of output-drivenness reconstructed in Magri (2017) thus subsumes Tesar's (2014) original theory as a special case.

References

Heinz, Jeffrey. 2005. Reconsidering linearity: Evidence from CV metathesis. In *Proceedings of WCCFL 24*, ed. John Alderete, Chung-hye Han, and Alexei Kochetov, 200–208. Somerville, MA, USA: Cascadilla Press.

de Lacy, Paul. 2006. *Markedness. reduction and preservation in phonology*. Cambridge University Press.

Magri, Giorgio. 2017. Idempotency, output-drivenness and the faithfulness triangle inequality: some consequences of McCarthy's (2013) categoricity generalization. *Journal of Logic, Language, and Information* URL `https://doi.org/10.1007/s10849-017-9256-0`.

Magri, Giorgio. 2018. Output-drivenness and partial phonological features. *Linguistic Inquiry* 49.

McCarthy, John J. 2003. OT constraints are categorical. *Phonology* 20:75–138.

McCarthy, John J., and Alan Prince. 1995. Faithfulness and reduplicative identity. In *University of massachusetts occasional papers in linguistics 18: Papers in optimality theory*, ed. Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey, 249–384. Amherst: GLSA.

Merchant, Nazarré. 2008. Discovering underlying forms: contrast pairs and ranking. Doctoral Dissertation, Rutgers University.

Moreton, Elliott. 2004. Non-computable functions in Optimality Theory. In *Optimality theory in phonology: A reader*, ed. John J. McCarthy, 141–163. Malden: MA: Wiley-Blackwell.

Rudin, Walter. 1953. *Principles of mathematical analysis*. McGraw-Hill Book Company.

Tesar, Bruce. 2014. *Output-driven phonology: Theory and learning*. Cambridge Studies in Linguistics.

Appendix: Proof of proposition 1

This appendix provides a proof of proposition 1, repeated below. Subsection A.1 constructs the correspondence relation $\rho_{\mathsf{a},\mathsf{b}}$ which appears in the second term on the right hand side of the identity (49). Subsections A.2, A.3, and A.4 then establish this identity (49) separately for the three faithfulness constraints Ident, Dep, and Max, respectively. As anticipated, the claim that Max satisfies the identity (49) can be straightforwardly derived from the claim that Dep satisfies it, because the two constraints are one the *inverse* of the other. The proof presented here simply reboots various ingredients of the extremely sophisticated analysis developed in Tesar (2014, chapter 3). This connection with Tesar's work is clarified in subsection A.5.

Proposition 1 *Consider two candidates* $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ *and* $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ *sharing the surface form* x*. Assume that the correspondence relations* $\rho_{\mathsf{a},\mathsf{x}}$ *and* $\rho_{\mathsf{b},\mathsf{x}}$ *are both one-to-one. If* $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{sim} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$*, then there exists a one-to-one correspondence relation* $\rho_{\mathsf{a},\mathsf{b}}$ *between the two strings* a *and* b *such that the following identity*

$$(49)\quad F\big(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}\big) = F\big(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}}\big) + F\big(\mathsf{a}, \mathsf{b}, \rho_{\mathsf{a},\mathsf{b}}\big)$$

holds when F *is* Max, Dep, *or* Ident$_\varphi$ *for any feature* φ *in the feature set* Φ. □

By Tesar's definition 5, the assumption $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ in proposition 1 means that each insertion, deletion, and featural disparity of the more similar candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ admits a corresponding identical disparity in the less similar candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$. For insertion and deletion disparities, this requirement is equivalent to the insertion and deletion clauses below. For featural disparities, this requirement is equivalent to the featural clause below because of the additional assumption that the two correspondence relations $\rho_{\mathsf{a},\mathsf{x}}$ and $\rho_{\mathsf{b},\mathsf{x}}$ are one-to-one.

insertion clause: Every segment of x which is epenthetic relative to the candidate $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$, is also epenthetic relative to the candidate $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$.

deletion clause: There exists an injective function $i_{\mathsf{b},\mathsf{a}}$ from the segments of b deleted relative to $(\mathsf{b}, \mathsf{x}, \rho_{\mathsf{b},\mathsf{x}})$ to the segments of a deleted relative to $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$ such that any two deleted segments corresponding through $i_{\mathsf{b},\mathsf{a}}$ agree on the value of every feature in Φ.

featural clause: For every feature φ in Φ, every segment x of x, every segment b of b, if $(\mathsf{b}, \mathsf{x}) \in \rho_{\mathsf{b},\mathsf{x}}$ and $\varphi(\mathsf{b}) \neq \varphi(\mathsf{x})$, there exists a segment a of a such that $(\mathsf{a}, \mathsf{x}) \in \rho_{\mathsf{a},\mathsf{x}}$ and $\varphi(\mathsf{a}) = \varphi(\mathsf{b})$.

A.1. The correspondence relation $\rho_{\mathsf{a},\mathsf{b}}$

The following lemma 1 constructs the correspondence relation $\rho_{\mathsf{a},\mathsf{b}}$ between the two strings a and b and shows that it is one-to-one. The definition (51) can be unpacked as follows: $\rho_{\mathsf{a},\mathsf{b}}$ is the union of the inverse $i^{-1}_{\mathsf{b},\mathsf{a}}$ of the injection $i_{\mathsf{b},\mathsf{a}}$ with the composition $\rho_{\mathsf{a},\mathsf{x}} \circ \rho^{-1}_{\mathsf{b},\mathsf{x}}$ of the correspondence relation $\rho_{\mathsf{a},\mathsf{x}}$ with the inverse of the correspondence relation $\rho_{\mathsf{b},\mathsf{x}}$.

(50)
$$\begin{tikzcd}
\mathsf{a} \arrow[r, "\rho_{\mathbf{a},\mathbf{x}}"] \arrow[dr, bend right, "\rho_{\mathbf{a},\mathbf{b}} = i^{-1}_{\mathbf{b},\mathbf{a}} \cup \left(\rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}\right)"'] & \mathsf{x} \arrow[d, "\rho^{-1}_{\mathbf{b},\mathbf{x}}"] \\
& \mathsf{b}
\end{tikzcd}$$

The lemma only uses a weak version of the DELETION CLAUSE, which does not require the deleted segments of $\mathbf{b}$ and $\mathbf{a}$ corresponding through the injection $i_{\mathbf{b},\mathbf{a}}$ to match relative to the features in the feature set Φ.

Lemma 1. *Consider two candidates $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ and $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ sharing the surface form $\boldsymbol{x}$. Assume that the correspondence relations $\rho_{\boldsymbol{a},\boldsymbol{x}}$ and $\rho_{\boldsymbol{b},\boldsymbol{x}}$ are both one-to-one and that:*

WEAK DELETION CLAUSE: *There exists an injective function $i_{\boldsymbol{b},\boldsymbol{a}}$ from the segments of $\boldsymbol{b}$ deleted relative to $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ to the segments of $\boldsymbol{a}$ deleted relative to $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$.*

The correspondence relation $\rho_{\boldsymbol{a},\boldsymbol{b}}$ between the two strings $\boldsymbol{a}$ and $\boldsymbol{b}$ defined by

(51) $\rho_{\boldsymbol{a},\boldsymbol{b}} = i^{-1}_{\boldsymbol{b},\boldsymbol{a}} \cup (\rho_{\boldsymbol{a},\boldsymbol{x}} \circ \rho^{-1}_{\boldsymbol{b},\boldsymbol{x}})$

is one-to-one. □

Proof. The relation $i^{-1}_{\mathbf{b},\mathbf{a}}$ is one-to-one because it is the inverse of the injective function $i_{\mathbf{b},\mathbf{a}}$. The relation $\rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}$ is one-to-one because it is the composition of two relations which are both one-to-one. Furthermore, the two relations $i^{-1}_{\mathbf{b},\mathbf{a}}$ and $\rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}$ have disjoint domains: the domain of the former is a subset of the set of segments of $\mathbf{a}$ which are deleted relative to $\rho_{\mathbf{a},\mathbf{x}}$; the domain of the latter is a subset of the set of segments of $\mathbf{a}$ which have a correspondent relative to $\rho_{\mathbf{a},\mathbf{x}}$. Analogously, the two relations $i^{-1}_{\mathbf{b},\mathbf{a}}$ and $\rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}$ have disjoint ranges. In conclusion, the relation $\rho_{\mathbf{a},\mathbf{b}}$ is one-to-one, because it is the union of two relations which are both one-to-one and have disjoint domains and ranges. □

A.2. Ident

The following lemma shows that the identity (49) holds for the faithfulness constraint $F =$ IDENT.

Lemma 2. *Consider two candidates $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ and $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ sharing the surface form $\boldsymbol{x}$. Assume that the correspondence relations $\rho_{\boldsymbol{a},\boldsymbol{x}}$ and $\rho_{\boldsymbol{b},\boldsymbol{x}}$ are both one-to-one and that:*

INSERTION CLAUSE: *Every segment which is epenthetic relative to the candidate $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$, is also epenthetic relative to the candidate $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$.*

DELETION CLAUSE: *There exists an injective function $i_{\boldsymbol{b},\boldsymbol{a}}$ from the segments of $\boldsymbol{b}$ deleted relative to $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ to the segments of $\boldsymbol{a}$ deleted relative to $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ such that any two deleted segments corresponding through $i_{\boldsymbol{b},\boldsymbol{a}}$ agree on the value of every feature in Φ.*

FEATURAL CLAUSE: *For every feature φ in Φ, every segment x of $\boldsymbol{x}$, every segment b of $\boldsymbol{b}$ such that $(b, x) \in \rho_{\boldsymbol{b},\boldsymbol{x}}$ and $\varphi(b) \neq \varphi(x)$, there exists a segment a such that $(a, x) \in \rho_{\boldsymbol{a},\boldsymbol{x}}$ and $\varphi(a) = \varphi(b)$.*

The faithfulness constraint $F =$ IDENT$_\varphi$ corresponding to any feature φ in Φ satisfies the identity (49) when the correspondence relation $\rho_{\boldsymbol{a},\boldsymbol{b}}$ is defined as in (51). □

Proof. Consider an arbitrary feature φ in the feature set Φ. A feature mismatch relative to the correspondence relation $\rho_{\mathbf{a},\mathbf{b}}$ defined in (51) can be characterized as follows in terms of the correspondence relations $\rho_{\mathbf{a},\mathbf{x}}$ and $\rho_{\mathbf{b},\mathbf{x}}$:

(52) $(\mathsf{a}, \mathsf{b}) \in \rho_{\mathbf{a},\mathbf{b}}$ and $\varphi(\mathsf{a}) \neq \varphi(\mathsf{b}) \iff$ there exists x such that $(\mathsf{a}, \mathsf{x}) \in \rho_{\mathbf{a},\mathbf{x}}$, $(\mathsf{b}, \mathsf{x}) \in \rho_{\mathbf{b},\mathbf{x}}$, $\varphi(\mathsf{a}) \neq \varphi(\mathsf{x})$, and $\varphi(\mathsf{b}) = \varphi(\mathsf{x})$.

The implication $\Longleftarrow$ trivially holds. Let me show that the reverse implication $\Longrightarrow$ holds as well. In fact, suppose that $(\mathsf{a}, \mathsf{b}) \in \rho_{\mathbf{a},\mathbf{b}}$ and $\varphi(\mathsf{a}) \neq \varphi(\mathsf{b})$. By definition (51), $\rho_{\mathbf{a},\mathbf{b}}$ is the union of $i^{-1}_{\mathbf{b},\mathbf{a}}$ and $\rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}$. The assumption $(\mathsf{a}, \mathsf{b}) \in \rho_{\mathbf{a},\mathbf{b}}$ thus means that either $(\mathsf{a}, \mathsf{b}) \in i^{-1}_{\mathbf{b},\mathbf{a}}$ or else $(\mathsf{a}, \mathsf{b}) \in \rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}$.

The former option $(\mathsf{a},\mathsf{b}) \in i^{-1}_{\mathbf{b},\mathbf{a}}$ is nonetheless impossible, because $\varphi(\mathsf{a}) \neq \varphi(\mathsf{b})$ while the DELETION CLAUSE requires $i_{\mathbf{b},\mathbf{a}}$ to only hold between segments which match for the value of every feature φ in Φ. Hence, it must be $(\mathsf{a},\mathsf{b}) \in \rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}$. This means in turn that there exists a segment x of $\mathbf{x}$ such that $(\mathsf{a},\mathsf{x}) \in \rho_{\mathbf{a},\mathbf{x}}$ and $(\mathsf{b},\mathsf{x}) \in \rho_{\mathbf{b},\mathbf{x}}$. If it were $\varphi(\mathsf{b}) \neq \varphi(\mathsf{x})$, the FEATURAL CLAUSE would then require $\varphi(\mathsf{a}) = \varphi(\mathsf{b})$, contradicting the assumption $\varphi(\mathsf{a}) \neq \varphi(\mathsf{b})$. Thus, it must be $\varphi(\mathsf{b}) = \varphi(\mathsf{x})$. The assumption $\varphi(\mathsf{a}) \neq \varphi(\mathsf{b})$ thus yields $\varphi(\mathsf{a}) \neq \varphi(\mathsf{x})$.

The number of violations assigned by the faithfulness constraint $\textsc{Ident}_\varphi$ to the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ can be expressed as in (53a). In step (53b), I have used the assumption that the correspondence relation $\rho_{\mathbf{a},\mathbf{x}}$ is one-to-one. In step (53c), I have added the condition "x has a $\rho_{\mathbf{b},\mathbf{x}}$-correspondent b". This condition is automatically satisfied because the INSERTION CLAUSE ensures that, if x had no $\rho_{\mathbf{b},\mathbf{x}}$-correspondent, then it would have no $\rho_{\mathbf{a},\mathbf{x}}$-correspondent. Because of the assumption that $\rho_{\mathbf{b},\mathbf{x}}$ is one-to-one, the $\rho_{\mathbf{b},\mathbf{x}}$-correspondent b of x is unique. In step (53d), I have split the two cases where b and x match and do not match relative to the feature φ.

(53)
$$
\begin{aligned}
&\textsc{Ident}_\varphi(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}) \\
&\overset{(a)}{=} \#\{\, (\mathsf{a},\mathsf{x}) \in \rho_{\mathbf{a},\mathbf{x}} \,|\, \varphi(\mathsf{a}) \neq \varphi(\mathsf{x}) \,\} \\
&\overset{(b)}{=} \#\{\, \mathsf{x} \,|\, \mathsf{x} \text{ has a } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondent } \mathsf{a} \text{ such that } \varphi(\mathsf{a}) \neq \varphi(\mathsf{x}) \,\} \\
&\overset{(c)}{=} \#\left\{ \mathsf{x} \,\middle|\, \begin{array}{l} \mathsf{x} \text{ has a } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondent } \mathsf{a} \text{ such that } \varphi(\mathsf{a}) \neq \varphi(\mathsf{x}) \text{ and} \\ \mathsf{x} \text{ has a } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondent } \mathsf{b} \end{array} \right\} \\
&\overset{(d)}{=} \underbrace{\#\left\{ \mathsf{x} \,\middle|\, \begin{array}{l} \mathsf{x} \text{ has a } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondent } \mathsf{a} \text{ such that } \varphi(\mathsf{a}) \neq \varphi(\mathsf{x}) \text{ and} \\ \mathsf{x} \text{ has a } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondent } \mathsf{b} \text{ such that } \varphi(\mathsf{b}) \neq \varphi(\mathsf{x}) \end{array} \right\}}_{\text{term I}} + \\
&\qquad + \underbrace{\#\left\{ \mathsf{x} \,\middle|\, \begin{array}{l} \mathsf{x} \text{ has a } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondent } \mathsf{a} \text{ such that } \varphi(\mathsf{a}) \neq \varphi(\mathsf{x}) \text{ and} \\ \mathsf{x} \text{ has a } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondent such that } \varphi(\mathsf{b}) = \varphi(\mathsf{x}) \end{array} \right\}}_{\text{term II}}
\end{aligned}
$$

Term I of (53) can be unpacked as in (54). In step (54a), I have dropped the condition that "x has a $\rho_{\mathbf{a},\mathbf{x}}$-correspondent a such that $\varphi(\mathsf{a}) \neq \varphi(\mathsf{x})$", because this condition is entailed by the condition that "x has a $\rho_{\mathbf{b},\mathbf{x}}$-correspondent b such that $\varphi(\mathsf{b}) \neq \varphi(\mathsf{x})$" plus the FEATURAL CLAUSE whereby $\varphi(\mathsf{b}) = \varphi(\mathsf{a})$ whenever $\varphi(\mathsf{b}) \neq \varphi(\mathsf{x})$. In step (54b), I have used the assumption that the correspondence relation $\rho_{\mathbf{b},\mathbf{x}}$ is one-to-one.

(54)
$$
\begin{aligned}
\text{term I} &= \#\left\{ \mathsf{x} \,\middle|\, \begin{array}{l} \mathsf{x} \text{ has a } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondent } \mathsf{a} \text{ such that } \varphi(\mathsf{a}) \neq \varphi(\mathsf{x}) \text{ and} \\ \mathsf{x} \text{ has a } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondent } \mathsf{b} \text{ such that } \varphi(\mathsf{b}) \neq \varphi(\mathsf{x}) \end{array} \right\} \\
&\overset{(a)}{=} \#\left\{ \mathsf{x} \,\middle|\, \mathsf{x} \text{ has a } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondent } \mathsf{b} \text{ such that } \varphi(\mathsf{b}) \neq \varphi(\mathsf{x}) \right\} \\
&\overset{(b)}{=} \#\{\, (\mathsf{b},\mathsf{x}) \in \rho_{\mathbf{b},\mathbf{x}} \,|\, \varphi(\mathsf{b}) \neq \varphi(\mathsf{x}) \,\} \\
&= \textsc{Ident}_\varphi(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})
\end{aligned}
$$

Term II of (53) can be unpacked as in (55). In step (55a), I have used the hypothesis that the correspondence relations $\rho_{\mathbf{a},\mathbf{x}}$ and $\rho_{\mathbf{b},\mathbf{x}}$ are both one-to-one. In step (55b), I have used the equivalence established in (52).

(55)
$$
\begin{aligned}
\text{term II} &= \#\left\{ \mathsf{x} \,\middle|\, \begin{array}{l} \mathsf{x} \text{ has a } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondent } \mathsf{a} \text{ such that } \varphi(\mathsf{a}) \neq \varphi(\mathsf{x}) \text{ and} \\ \mathsf{x} \text{ has a } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondent } \mathsf{b} \text{ such that } \varphi(\mathsf{b}) = \varphi(\mathsf{x}) \end{array} \right\} \\
&\overset{(a)}{=} \#\left\{ (\mathsf{a},\mathsf{b}) \,\middle|\, \begin{array}{l} \text{there exists } \mathsf{x} \text{ such that } (\mathsf{a},\mathsf{x}) \in \rho_{\mathbf{a},\mathbf{x}},\ (\mathsf{b},\mathsf{x}) \in \rho_{\mathbf{b},\mathbf{x}}, \\ \varphi(\mathsf{a}) \neq \varphi(\mathsf{x}), \text{ and } \varphi(\mathsf{b}) = \varphi(\mathsf{x}) \end{array} \right\} \\
&\overset{(b)}{=} \#\{ (\mathsf{a},\mathsf{b}) |\ (\mathsf{a},\mathsf{b}) \in \rho_{\mathbf{a},\mathbf{b}} \text{ and } \varphi(\mathsf{a}) \neq \varphi(\mathsf{b}) \} \\
&= \textsc{Ident}_\varphi(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})
\end{aligned}
$$

The inequality (49) for $F = \textsc{Ident}_\varphi$ follows from (53), (54), and (55). □

If the INSERTION CLAUSE is dropped from the assumptions of lemma 2, the preceding proof still holds, with the only modification that "=" needs to be replaced by "≥" in step (53b). In the end, I thus conclude that $F = \text{IDENT}_{\varphi}$ does not satisfy the identity (49) but rather the corresponding inequality $F\big(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}\big) \geq F\big(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}}\big) + F\big(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}}\big)$. This conclusion still suffices to establish the desired relationship between $\leq^{\Phi}_{\text{sim}}$ and $\leq^{\mathcal{F}}_{\text{sim}}$ because the definition 6 of $\leq^{\mathcal{F}}_{\text{sim}}$ indeed only requires the inequality to hold, not the identity.

A.3. DEP

The following lemma shows that the identity (49) holds for the faithfulness constraint $F = \text{DEP}$. The lemma does not require the FEATURAL CLAUSE. Furthermore, it uses only a weaker version of the DELETION CLAUSE, which does not require the deleted segments which correspond through the injection $i_{\mathbf{b},\mathbf{a}}$ to agree on every feature.

Lemma 3. *Consider two candidates $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ and $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ sharing the surface form $\mathbf{x}$. Assume that the correspondence relations $\rho_{\mathbf{a},\mathbf{x}}$ and $\rho_{\mathbf{b},\mathbf{x}}$ are both one-to-one and that:*

INSERTION CLAUSE: *Every segment x which is epenthetic relative to the candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$, is also epenthetic relative to the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$.*

WEAK DELETION CLAUSE: *There exists an injective function $i_{\mathbf{b},\mathbf{a}}$ from the segments of $\mathbf{b}$ deleted relative to $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ to the segments of $\mathbf{a}$ deleted relative to $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$.*

The faithfulness constraint $F = \text{DEP}$ satisfies the identity (49) when the correspondence relation $\rho_{\mathbf{a},\mathbf{b}}$ is defined as in (51). □

Proof. An insertion relative to the correspondence relation $\rho_{\mathbf{a},\mathbf{b}}$ defined in (51) can be characterized as follows in terms of the two correspondence relations $\rho_{\mathbf{a},\mathbf{x}}$ and $\rho_{\mathbf{b},\mathbf{x}}$:

(56) b has no $\rho_{\mathbf{a},\mathbf{b}}$-correspondent $\Longleftrightarrow$ b has a $\rho_{\mathbf{b},\mathbf{x}}$-correspondent x which in turn has no $\rho_{\mathbf{a},\mathbf{x}}$-correspondents.

Let me prove the implication $\Longrightarrow$. Assume that a segment b of $\mathbf{b}$ has no correspondent segment in $\mathbf{a}$ according to $\rho_{\mathbf{a},\mathbf{b}}$. By virtue of the definition (51) of the relation $\rho_{\mathbf{a},\mathbf{b}}$, this means that b has neither a correspondent relative to $i^{-1}_{\mathbf{b},\mathbf{a}}$ nor a correspondent relative to $\rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}$. The fact that b has no correspondent relative to $i^{-1}_{\mathbf{b},\mathbf{a}}$ means that b is not deleted but rather has a correspondent x relative to $\rho_{\mathbf{b},\mathbf{x}}$. The fact that b has no correspondent relative to $\rho_{\mathbf{a},\mathbf{x}} \circ \rho^{-1}_{\mathbf{b},\mathbf{x}}$ then means that this segment x has no correspondent relative to $\rho_{\mathbf{a},\mathbf{x}}$. An analogous reasoning proves the reverse implication $\Longleftarrow$ in (56).

The number of violations assigned by DEP to the candidate $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ can be expressed as in (57a). In (57b), I have split two cases depending on whether the segment x admits or not a corresponding segment according to $\rho_{\mathbf{b},\mathbf{x}}$.

(57)
$$\begin{aligned}
&\text{DEP}\big(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}}\big)\\
&\overset{(a)}{=} \#\big\{\, \mathsf{x} \,\big|\, \mathsf{x} \text{ has no } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondents} \,\big\}\\
&\overset{(b)}{=} \underbrace{\#\left\{ \mathsf{x} \,\middle|\, \begin{array}{l} \mathsf{x} \text{ has no } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondents} \\ \mathsf{x} \text{ has no } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondents} \end{array} \right\}}_{\text{term I}} + \underbrace{\#\left\{ \mathsf{x} \,\middle|\, \begin{array}{l} \mathsf{x} \text{ has no } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondents} \\ \mathsf{x} \text{ has a } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondent } \mathsf{b} \end{array} \right\}}_{\text{term II}}
\end{aligned}$$

Term I of (57) can be unpacked as in (58). The clause "x has no $\rho_{\mathbf{a},\mathbf{x}}$-correspondents" can be dropped in (58a) because it is entailed by the clause "x has no $\rho_{\mathbf{b},\mathbf{x}}$-correspondents" by virtue of the INSERTION CLAUSE, which ensures that x is epenthetic relative to $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ whenever it is epenthetic relative to $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$.

(58)
$$\begin{aligned}
\text{term I} &= \#\left\{ \mathsf{x} \,\middle|\, \begin{array}{l} \mathsf{x} \text{ has no } \rho_{\mathbf{a},\mathbf{x}}\text{-correspondents} \\ \mathsf{x} \text{ has no } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondents} \end{array} \right\}\\
&\overset{(a)}{=} \#\big\{\mathsf{x} \,\big|\, \mathsf{x} \text{ has no } \rho_{\mathbf{b},\mathbf{x}}\text{-correspondents}\big\}\\
&= \text{DEP}\big(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}}\big)
\end{aligned}$$

Term II of (57) can be unpacked as in (59). In (59a), I have used the hypothesis that all correspondence relations are one-to-one. In (59b), I have used the equivalence (56).

(59)
$$\begin{aligned}
\text{term II} &= \#\left\{\mathsf{x}\;\middle|\;\begin{array}{l}\mathsf{x}\text{ has no }\rho_{\mathbf{a},\mathbf{x}}\text{-correspondents}\\ \mathsf{x}\text{ has a }\rho_{\mathbf{b},\mathbf{x}}\text{-correspondent }\mathsf{b}\end{array}\right\}\\
&\overset{(a)}{=} \#\big\{\mathsf{b}\;\big|\;\mathsf{b}\text{ has a }\rho_{\mathbf{b},\mathbf{x}}\text{-correspondent }\mathsf{x}\text{ which has no }\rho_{\mathbf{a},\mathbf{x}}\text{-correspondents}\big\}\\
&\overset{(b)}{=} \#\big\{\mathsf{b}\;\big|\;\mathsf{b}\text{ has no }\rho_{\mathbf{a},\mathbf{b}}\text{-correspondents}\big\}\\
&= \textsc{Dep}\big(\mathbf{a},\mathbf{b},\rho_{\mathbf{a},\mathbf{b}}\big)
\end{aligned}$$

The identity (49) for $F = \textsc{Dep}$ follows from (57), (58), and (59). □

A.4. Max

The following lemma is identical to the preceding lemma 3, apart from the fact that it looks at Max instead of Dep. The lemma could indeed be proven in exactly the same way lemma 4 has been proved in the preceding section. Here, I use instead a simpler strategy, whereby the result for Max is deduced straightforwardly from the result for Dep, by exploiting the fact that they are one the *inverse* of the other.

Lemma 4. *Consider two candidates $(\mathbf{a},\mathbf{x},\rho_{\mathbf{a},\mathbf{x}})$ and $(\mathbf{b},\mathbf{x},\rho_{\mathbf{b},\mathbf{x}})$ sharing the surface form $\mathbf{x}$. Assume that the correspondence relations $\rho_{\mathbf{a},\mathbf{x}}$ and $\rho_{\mathbf{b},\mathbf{x}}$ are both one-to-one and that:*

INSERTION CLAUSE: *Every segment x which is epenthetic relative to the candidate $(\mathbf{b},\mathbf{x},\rho_{\mathbf{b},\mathbf{x}})$, is also epenthetic relative to the candidate $(\mathbf{a},\mathbf{x},\rho_{\mathbf{a},\mathbf{x}})$.*

WEAK DELETION CLAUSE: *There exists an injective function $i_{\mathbf{b},\mathbf{a}}$ from the segments of $\mathbf{b}$ deleted relative to $(\mathbf{b},\mathbf{x},\rho_{\mathbf{b},\mathbf{x}})$ to the segments of $\mathbf{a}$ deleted relative to $(\mathbf{a},\mathbf{x},\rho_{\mathbf{a},\mathbf{x}})$.*

The faithfulness constraint $F = \textsc{Max}$ satisfies the identity (49) when the correspondence relation $\rho_{\mathbf{a},\mathbf{b}}$ is defined as in (51). □

Proof. A deletion relative to the relation $\rho_{\mathbf{a},\mathbf{b}}$ defined in (51) can be characterized as follows $\rho_{\mathbf{a},\mathbf{b}}$ in terms of the correspondence relation $\rho_{\mathbf{a},\mathbf{x}}$.

(60) a has no $\rho_{\mathbf{a},\mathbf{b}}$-correspondent $\Longrightarrow$ a has no $\rho_{\mathbf{a},\mathbf{x}}$-correspondent.

In fact, suppose by contradiction that a segment a of $\mathbf{a}$ has no corresponding segment in $\mathbf{b}$ relative to $\rho_{\mathbf{a},\mathbf{b}}$ but that a has a corresponding segment x in $\mathbf{x}$ relative to $\rho_{\mathbf{a},\mathbf{x}}$. Thus, x must have a corresponding segment in $\mathbf{b}$ relative to $\rho_{\mathbf{b},\mathbf{x}}$, because of the INSERTION CLAUSE ensures that x is epenthetic relative to $\rho_{\mathbf{a},\mathbf{x}}$ whenever it is epenthetic relative to $\rho_{\mathbf{b},\mathbf{x}}$. The conclusion that a has a correspondent x relative to $\rho_{\mathbf{a},\mathbf{x}}$ which in turn has a correspondent b relative to $\rho_{\mathbf{b},\mathbf{x}}$ contradicts the hypothesis that a has no correspondent relative to $\rho_{\mathbf{a},\mathbf{b}}$.

To simplify the notation throughout the rest of the proof, I adopt the positions $R_{\mathbf{x},\mathbf{a}} = \rho^{-1}_{\mathbf{a},\mathbf{x}}$ and $R_{\mathbf{b},\mathbf{a}} = \rho^{-1}_{\mathbf{a},\mathbf{b}}$. The two candidates $(\mathbf{x},\mathbf{a},R_{\mathbf{x},\mathbf{a}})$ and $(\mathbf{b},\mathbf{a},R_{\mathbf{b},\mathbf{a}})$ are therefore the inverse of the two candidates $(\mathbf{a},\mathbf{x},\rho_{\mathbf{a},\mathbf{x}})$ and $(\mathbf{a},\mathbf{b},\rho_{\mathbf{a},\mathbf{b}})$, respectively. Condition (60) can be restated as in (61) in terms of these two inverse candidates $(\mathbf{x},\mathbf{a},R_{\mathbf{x},\mathbf{a}})$ and $(\mathbf{b},\mathbf{a},R_{\mathbf{b},\mathbf{a}})$. Furthermore, the condition (56) established in the preceding proof of lemma 3 says that the correspondence relation $\rho_{\mathbf{b},\mathbf{x}}$ maps the segments of $\mathbf{b}$ deleted relative to $(\mathbf{b},\mathbf{a},R_{\mathbf{b},\mathbf{a}})$ to the segments of $\mathbf{x}$ deleted relative to $(\mathbf{x},\mathbf{a},R_{\mathbf{x},\mathbf{a}})$. Since $\rho_{\mathbf{b},\mathbf{x}}$ is one-to-one, then it is in particular injective. In conclusion, condition (56) can be restated in terms of the two inverse candidates $(\mathbf{x},\mathbf{a},R_{\mathbf{x},\mathbf{a}})$ and $(\mathbf{b},\mathbf{a},R_{\mathbf{b},\mathbf{a}})$ as in (62).

(61) Every segment a of $\mathbf{a}$ which is epenthetic relative to the candidate $(\mathbf{b},\mathbf{a},R_{\mathbf{b},\mathbf{a}})$ is also epenthetic relative to the candidate $(\mathbf{x},\mathbf{a},R_{\mathbf{x},\mathbf{a}})$.

(62) $\rho_{\mathbf{b},\mathbf{x}}$ is an injection from the segments of $\mathbf{b}$ deleted relative to $(\mathbf{b},\mathbf{a},R_{\mathbf{b},\mathbf{a}})$ to the segments of $\mathbf{x}$ deleted relative to $(\mathbf{x},\mathbf{a},R_{\mathbf{x},\mathbf{a}})$.

Define the correspondence relation $R_{\mathbf{x},\mathbf{b}}$ between the two strings $\mathbf{x}$ and $\mathbf{b}$ as $R_{\mathbf{x},\mathbf{b}} = \rho^{-1}_{\mathbf{b},\mathbf{x}} \cup (R_{\mathbf{x},\mathbf{a}} \circ R^{-1}_{\mathbf{b},\mathbf{a}})$, namely as the union of the inverse of the injection $\rho_{\mathbf{b},\mathbf{x}}$ with the composition $R_{\mathbf{x},\mathbf{a}} \circ R^{-1}_{\mathbf{b},\mathbf{a}}$ of the correspondence relation $R_{\mathbf{x},\mathbf{a}}$ with the inverse of the correspondence relation $R_{\mathbf{b},\mathbf{a}}$, as depicted in (63).

(63)
$$\begin{tikzcd} \mathsf{x} \arrow[r, "R_{\mathsf{x},\mathsf{a}}"] \arrow[dr, bend right, "R_{\mathsf{x},\mathsf{b}} = \rho^{-1}_{\mathsf{b},\mathsf{x}} \cup \left(R_{\mathsf{x},\mathsf{a}} \circ R^{-1}_{\mathsf{b},\mathsf{a}}\right)"'] & \mathsf{a} \arrow[d, "R^{-1}_{\mathsf{b},\mathsf{a}}"] \\ & \mathsf{b} \end{tikzcd}$$

The two conditions (61) and (62) say that the INSERTION CLAUSE and the WEAK DELETION CLAUSE of lemma 3 hold for the two candidates $(\mathsf{x}, \mathsf{a}, R_{\mathsf{x},\mathsf{a}})$ and $(\mathsf{b}, \mathsf{a}, R_{\mathsf{b},\mathsf{a}})$. Furthermore, the definition (63) of the correspondence relation $R_{\mathsf{x},\mathsf{b}}$ is formally analogous to the definition (51) of the correspondence relation $\rho_{\mathsf{a},\mathsf{b}}$. Lemma 3 proven above thus ensures that the faithfulness constraint DEP satisfies the following identity (64).

(64) $\textsc{Dep}(\mathsf{x}, \mathsf{a}, R_{\mathsf{x},\mathsf{a}}) = \textsc{Dep}(\mathsf{b}, \mathsf{a}, R_{\mathsf{b},\mathsf{a}}) + \textsc{Dep}(\mathsf{x}, \mathsf{b}, R_{\mathsf{x},\mathsf{b}})$

Recall that MAX and DEP are one the inverse of the other in the sense of the identity (16) discussed in subsection 2.1. The identity (64) between DEP-violations thus yields the corresponding identity (65) between the MAX violations of the corresponding inverse candidates.

(65) $\textsc{Max}\big(\mathsf{a}, \mathsf{x}, R^{-1}_{\mathsf{x},\mathsf{a}}\big) = \textsc{Max}\big(\mathsf{a}, \mathsf{b}, R^{-1}_{\mathsf{b},\mathsf{a}}\big) + \textsc{Max}\big(\mathsf{b}, \mathsf{x}, R^{-1}_{\mathsf{x},\mathsf{b}}\big)$

The latter identity (65) coincides with the target identity (49) for $F = \textsc{Max}$ because of the positions $R_{\mathsf{x},\mathsf{a}} = \rho^{-1}_{\mathsf{a},\mathsf{x}}$ and $R_{\mathsf{b},\mathsf{a}} = \rho^{-1}_{\mathsf{a},\mathsf{b}}$ together with the identity $R_{\mathsf{x},\mathsf{b}} = \rho^{-1}_{\mathsf{b},\mathsf{x}}$ established in (66).

(66)
$$\begin{aligned} R_{\mathsf{x},\mathsf{b}} &\overset{(a)}{=} \rho^{-1}_{\mathsf{b},\mathsf{x}} \cup \big(R_{\mathsf{x},\mathsf{a}} \circ R^{-1}_{\mathsf{b},\mathsf{a}}\big) \\ &\overset{(b)}{=} \rho^{-1}_{\mathsf{b},\mathsf{x}} \cup \big(\rho^{-1}_{\mathsf{a},\mathsf{x}} \circ \rho_{\mathsf{a},\mathsf{b}}\big) \\ &\overset{(c)}{=} \rho^{-1}_{\mathsf{b},\mathsf{x}} \cup \Big[\rho^{-1}_{\mathsf{a},\mathsf{x}} \circ \Big(i^{-1}_{\mathsf{b},\mathsf{a}} \cup \big(\rho_{\mathsf{a},\mathsf{x}} \circ \rho^{-1}_{\mathsf{b},\mathsf{x}}\big)\Big)\Big] \\ &= \rho^{-1}_{\mathsf{b},\mathsf{x}} \cup \big(\rho^{-1}_{\mathsf{a},\mathsf{x}} \circ i^{-1}_{\mathsf{b},\mathsf{a}}\big) \cup \rho^{-1}_{\mathsf{b},\mathsf{x}} \\ &\overset{(d)}{=} \rho^{-1}_{\mathsf{b},\mathsf{x}} \end{aligned}$$

In step (66a), I have used the definition (63) of $R_{\mathsf{x},\mathsf{b}}$. In step (66b), I have used the positions $R_{\mathsf{x},\mathsf{a}} = \rho^{-1}_{\mathsf{a},\mathsf{x}}$ and $R_{\mathsf{b},\mathsf{a}} = \rho^{-1}_{\mathsf{a},\mathsf{b}}$. In step (66c), I have used the definition (51) of the correspondence relation $\rho_{\mathsf{a},\mathsf{b}}$. Finally in step (66d), I have used the fact that the composition $\rho^{-1}_{\mathsf{a},\mathsf{x}} \circ i^{-1}_{\mathsf{b},\mathsf{a}}$ is the empty relation: since the range of $i_{\mathsf{b},\mathsf{a}}$ is a subset of the deleted segments of $(\mathsf{a}, \mathsf{x}, \rho_{\mathsf{a},\mathsf{x}})$, the condition $(\mathsf{a}, \mathsf{b}) \in i^{-1}_{\mathsf{b},\mathsf{a}}$ means that a has no $\rho_{\mathsf{a},\mathsf{x}}$-correspondents. □

A.5. Connection with an intermediate result of Tesar (2013; chapter 3)

Lemmas 1-4 provide a self contained proof of proposition 1. As already mentioned, this proof essentially simply re-boots various ingredients of the very sophisticated analysis developed in Tesar (2014, chapter 3). This subsection makes the connection between proposition 1 and Tesar's analysis more explicit, by deriving (a weaker but still sufficient version of) proposition 1 in a snapshot from an involved intermediate result of Tesar's analysis stated below as lemma 5 (based on Tesar 2014, chapter 3, pp. 92-93).

Lemma 5. *Consider two candidates $(\boldsymbol{a}, \mathsf{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ and $(\boldsymbol{b}, \mathsf{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ sharing the surface form $\boldsymbol{x}$. Assume they satisfy Tesar's similarity inequality $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}}) \leq^{\Phi}_{sim} (\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$. For every other candidate $(\boldsymbol{b}, \boldsymbol{y}, \rho_{\boldsymbol{b},\boldsymbol{y}})$, it is possible to define a correspondence relation $\rho_{\boldsymbol{a},\boldsymbol{y}}$ between the two strings $\boldsymbol{a}$ and $\boldsymbol{y}$ such that:*

(67) "Every disparity of $(\boldsymbol{a}, \boldsymbol{y}, \rho_{\boldsymbol{a},\boldsymbol{y}})$ that lacks a corresponding disparity in $(\boldsymbol{b}, \boldsymbol{y}, \rho_{\boldsymbol{b},\boldsymbol{y}})$ has an analogous disparity in $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ that has no corresponding disparity in $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$." (p. 93)

where the notion of correspondence *between the disparities of $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ and $(\boldsymbol{b}, \boldsymbol{x}, \rho_{\boldsymbol{b},\boldsymbol{x}})$ and between the disparities of $(\boldsymbol{a}, \boldsymbol{y}, \rho_{\boldsymbol{a},\boldsymbol{y}})$ and $(\boldsymbol{b}, \boldsymbol{y}, \rho_{\boldsymbol{b},\boldsymbol{y}})$ is the one provided by Tesar's definition 3 above; and the notion of* analogy *between the disparities of $(\boldsymbol{a}, \boldsymbol{y}, \rho_{\boldsymbol{a},\boldsymbol{y}})$ and $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\boldsymbol{a},\boldsymbol{x}})$ is defined as follows:*

(a) *The insertion disparities relative to the candidate* $(\boldsymbol{a}, \boldsymbol{y}, \rho_{\mathbf{a},\mathbf{y}})$ *have* analogous *insertion disparities relative to the candidate* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\mathbf{a},\mathbf{x}})$ *provided for every segment* y *of* $\boldsymbol{y}$ *epenthetic relative to* $(\boldsymbol{a}, \boldsymbol{y}, \rho_{\mathbf{a},\mathbf{y}})$*, there exists a segment* x *of* $\boldsymbol{x}$ *epenthetic relative to* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\mathbf{a},\mathbf{x}})$ *such that there exists a segment* b *of* $\boldsymbol{b}$ *such that* $(\mathsf{b}, \mathsf{y}) \in \rho_{\mathbf{b},\mathbf{y}}$ *and* $(\mathsf{b}, \mathsf{x}) \in \rho_{\mathbf{b},\mathbf{x}}$*.*

(b) *The deletion disparities relative to the candidate* $(\boldsymbol{a}, \boldsymbol{y}, \rho_{\mathbf{a},\mathbf{y}})$ *have* analogous *deletion disparities relative to the candidate* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\mathbf{a},\mathbf{x}})$ *provided every segment of* $\boldsymbol{a}$ *deleted relative to the candidate* $(\boldsymbol{a}, \boldsymbol{y}, \rho_{\mathbf{a},\mathbf{y}})$ *is also deleted relative to the candidate* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\mathbf{a},\mathbf{x}})$*.*

(c) *For a feature* $\varphi \in \Phi$*, the* φ*-disparities relative to the candidate* $(\boldsymbol{a}, \boldsymbol{y}, \rho_{\mathbf{a},\mathbf{y}})$ *have* analogous φ*-disparities relative to the candidate* $(\boldsymbol{a}, \boldsymbol{x}, \rho_{\mathbf{a},\mathbf{x}})$ *provided for every pair* $(\mathsf{a}, \mathsf{y}) \in \rho_{\mathbf{a},\mathbf{y}}$ *such that* $\varphi(\mathsf{a}) \neq \varphi(\mathsf{y})$*, there exists a pair* $(\mathsf{a}, \mathsf{x}) \in \rho_{\mathbf{a},\mathbf{x}}$ *(involving the same underlying segment* a*) such that* $\varphi(\mathsf{a}) \neq \varphi(\mathsf{x})$*.* □

Assume that the candidate set satisfies the U-reflexivity axiom (22). The existence of the candidate $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ thus entails in particular the existence of the identity candidate $(\mathbf{b}, \mathbf{b}, \mathbb{I}_{\mathbf{b},\mathbf{b}})$ corresponding to the underlying string $\mathbf{b}$. Since Tesar's claim (67) holds for every candidate $(\mathbf{b}, \mathbf{y}, \rho_{\mathbf{b},\mathbf{y}})$, it holds in particular with the positions (68), namely it holds in particular when the candidate $(\mathbf{b}, \mathbf{y}, \rho_{\mathbf{b},\mathbf{y}})$ is the identity candidate $(\mathbf{b}, \mathbf{b}, \mathbb{I}_{\mathbf{b},\mathbf{b}})$.

(68) $\mathbf{y} = \mathbf{b}$, $\rho_{\mathbf{b},\mathbf{y}} = \mathbb{I}_{\mathbf{b},\mathbf{b}}$.

With these positions, the candidate $(\mathbf{a}, \mathbf{y}, \rho_{\mathbf{a},\mathbf{y}})$ becomes $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$. Since the identity candidate $(\mathbf{b}, \mathbf{y}, \rho_{\mathbf{b},\mathbf{y}}) = (\mathbf{b}, \mathbf{b}, \mathbb{I}_{\mathbf{b},\mathbf{b}})$ has no disparities, Tesar's claim (67) becomes:

(69) Every disparity of $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ has an analogous disparity in $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ that has no corresponding disparity in $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$.

For concreteness, let's focus for instance on insertion disparities. The insertion disparities of $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ can be sorted into two disjoint subsets, depending on whether they correspond to insertion disparities of $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$ or not, as stated in (70).

(70)

$$\left\{\begin{array}{c}\text{insertion}\\ \text{disparities of}\\ (\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})\end{array}\right\} = \left\{\begin{array}{c}\text{insertion}\\ \text{disparities of}\\ (\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})\\ \textbf{corresponding}\\ \textbf{to insertion}\\ \textbf{disparities of}\\ (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})\end{array}\right\} \cup \underset{\vdots \atop (*)}{\left\{\begin{array}{c}\text{insertion}\\ \text{disparities of}\\ (\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})\\ \textit{\textbf{not}}\ \textbf{corre-}\\ \textbf{sponding to}\\ \textbf{insertion}\\ \textbf{disparities of}\\ (\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})\end{array}\right\}}$$

Claim (69) now ensures that all the disparities of $(\mathbf{a}, \mathbf{x}, \rho_{\mathbf{a},\mathbf{x}})$ which are analogous to the disparities of $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ crucially belong to the set marked (*) in (70), namely they do not correspond to insertion disparities of $(\mathbf{b}, \mathbf{x}, \rho_{\mathbf{b},\mathbf{x}})$. In other words, the insertion disparities in the set (*) can be split into two disjoint subsets, depending on whether they are analogues to the insertion disparities of $(\mathbf{a}, \mathbf{b}, \rho_{\mathbf{a},\mathbf{b}})$ or not, yielding the further decomposition in (71).

(71)
$$\underbrace{\left\{\begin{array}{c}\text{insertion}\\ \text{disparities of}\\ (\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}})\end{array}\right\}}_{A} = \underbrace{\left\{\begin{array}{c}\text{insertion}\\ \text{disparities of}\\ (\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}})\\ \text{corresponding}\\ \text{to the insertion}\\ \text{disparities of}\\ (\mathsf{b},\mathsf{x},\rho_{\mathsf{b},\mathsf{x}})\end{array}\right\}}_{B} \cup \underbrace{\left\{\begin{array}{c}\text{insertion}\\ \text{disparities of}\\ (\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}})\\ \textit{not}\\ \text{corresponding}\\ \text{to insertion}\\ \text{disparities of}\\ (\mathsf{b},\mathsf{x},\rho_{\mathsf{b},\mathsf{x}})\\ \textbf{and}\\ \textbf{analogous to}\\ \textbf{the insertion}\\ \textbf{disparities of}\\ (\mathsf{a},\mathsf{b},\rho_{\mathsf{a},\mathsf{b}})\end{array}\right\}}_{C} \cup \left\{\begin{array}{c}\text{insertion}\\ \text{disparities of}\\ (\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}})\\ \textit{not}\\ \text{corresponding}\\ \text{to insertion}\\ \text{disparities of}\\ (\mathsf{b},\mathsf{x},\rho_{\mathsf{b},\mathsf{x}})\\ \textbf{and \textit{not}}\\ \textbf{analogous to}\\ \textbf{the insertion}\\ \textbf{disparities of}\\ (\mathsf{a},\mathsf{b},\rho_{\mathsf{a},\mathsf{b}})\end{array}\right\}$$

Since the three sets on the right hand side of (71) are by definition mutually disjoint, I obtain the inequality $|A| \geq |B| + |C|$ among the cardinalities of the three sets A, B, and C. The cardinality $|A|$ is the number $\textsc{Dep}(\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}})$ of violations assigned by Dep to the candidate $(\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}})$. The assumption $(\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}}) \leq^{\Phi}_{\text{sim}} (\mathsf{b},\mathsf{x},\rho_{\mathsf{b},\mathsf{x}})$ means in particular that every insertion disparity of $(\mathsf{b},\mathsf{x},\rho_{\mathsf{b},\mathsf{x}})$ has a corresponding insertion disparity in $(\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}})$. Since furthermore this correspondence between insertion disparities is one-to-one, the cardinality $|B|$ is the number $\textsc{Dep}(\mathsf{b},\mathsf{x},\rho_{\mathsf{b},\mathsf{x}})$ of violations assigned by Dep to the candidate $(\mathsf{b},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}})$. Finally, claim (69) ensures that every disparity of $(\mathsf{a},\mathsf{b},\rho_{\mathsf{a},\mathsf{b}})$ has an analogous disparity in the set C. Since furthermore this relation of analogy is one-to-one, the cardinality $|C|$ is the number $\textsc{Dep}(\mathsf{a},\mathsf{b},\rho_{\mathsf{a},\mathsf{b}})$ of violations assigned by Dep to the candidate $(\mathsf{a},\mathsf{b},\rho_{\mathsf{a},\mathsf{b}})$. In conclusion, the inequality $|A| \geq |B| + |C|$ yields the following inequality

(72) $$F(\mathsf{a},\mathsf{x},\rho_{\mathsf{a},\mathsf{x}}) \geq F(\mathsf{b},\mathsf{x},\rho_{\mathsf{b},\mathsf{x}}) + F(\mathsf{a},\mathsf{b},\rho_{\mathsf{a},\mathsf{b}})$$

for $F = \textsc{Dep}$. An analogous reasoning yields this same conclusion for $F = \textsc{Max}$ and $F = \textsc{Ident}_{\varphi}$ for every feature $\varphi \in \Phi$. This inequality (72) is weaker than the identity (49) ensured by proposition 1. Yet, the inequality still suffices to establish the desired relationship between the relations $\leq^{\Phi}_{\text{sim}}$ and $\leq^{\mathcal{F}}_{\text{sim}}$ because the definition 6 of $\leq^{\mathcal{F}}_{\text{sim}}$ indeed only requires the inequality to hold, not the identity.