(to appear in CLS 48 Proceedings)

# Paradigm learning and subanalysis complexity

Sebastian Bank and Jochen Trommer
University of Leipzig

## 1 Introduction

The algorithmic analysis of inflectional paradigms is a branch of research still in its infancy. Every formal description of inflectional systems requires on the one hand to assign meaning to invariant forms found in different paradigm cells, and on the other hand to identify invariant forms by segmenting paradigm cells into subanalyzed forms. While these two tasks are inherently intertwined, current approaches either assign meanings to paradigms of presegmented morphemes (Pertsova 2011) or identify morpheme boundaries solely on phonological grounds ignoring the meaning of morphemes (Harris 1955; Langer 1991; Goldsmith 2010; Saffran *et al.* 1996). On the other hand, theoretical morphologists assume sophisticated decompositions of paradigms where meaning assignment and subanalysis are optimized simultaneously and are thus meant to incorporate each other.

Consider for example the German verbal agreement paradigm in (1). A reasonably successfull application of the completely form-based approach to segmentation is Stemming, i.e. the identification of lexical stems in inflectional word forms (Lovins 1968; Porter 1980; Goldsmith 2010). Standard Stemming would assign the following basic segmentation, which separates the stem *glaub* 'to believe' from the suffixes expressing tense and subject agreement:

(1) **German present and past verbal agreement after stem identification**

a. PRESENT

| | SG | PL |
|---|---|---|
| 1 | glaub-**e** | glaub-**en** |
| 2 | glaub-**st** | glaub-**t** |
| 3 | glaub-**t** | glaub-**en** |

b. PAST

| | SG | PL |
|---|---|---|
| 1 | glaub-**te** | glaub-**ten** |
| 2 | glaub-**test** | glaub-**tet** |
| 3 | glaub-**te** | glaub-**ten** |

However, if we try to assign meanings to the resulting suffix strings it becomes obvious that they have to be further subanalyzed: They contain the separable past tense suffix *-te*, whose segmentation (2) reveals that most of the agreement markers appearing in the present tense are also used in the corresponding past forms.

(2) **German verbal agreement with minimal subanalysis**

a. PRESENT

| | SG | PL |
|---|---|---|
| 1 | glaub-**e** | glaub-**(e)n** |
| 2 | glaub-**st** | glaub-**t** |
| 3 | glaub-**t** | glaub-**(e)n** |

b. PAST

| | SG | PL |
|---|---|---|
| 1 | glaub-**te** | glaub-**te-n** |
| 2 | glaub-**te-st** | glaub-**te-t** |
| 3 | glaub-**te** | glaub-**te-n** |

While half of the suffix strings starting with *-te* can be discovered solely by their forms, the fact that they all occur in the past and lead to a natural division into tense

and agreement markers is only available if their (potential) meanings are taken into consideration.

In section 2, we introduce an incremental learning algorithm, which addresses this problem and performs combined morpheme segmentation and meaning assignment which is driven by the preference for maximally reliable (accurate) and general mappings between form and meaning. As the learner only optimizes in small, incremental steps, it does not explore the full search space of possible analyzes which becomes unfeasible even for small paradigms if one considers all analytical options (different possible segmentations, meaning assignments, homonymic morphemes). In fact, segmentation and homonymy are not even a genuine notion in the algorithm: they emerge from the successive removal of learned affixal material.

If we carefully compare the strings in (2), an exhaustive search for the globally optimal subanalysis considering each and every segmentation with every meaning assignment looks pretty unnatural: Because tense is zero-marked in the present (although one might argue that the 1SG suffix *-e* expresses also present tense) we can identify the person/number markers without further segmentation. Similarly 1SG is zero in past tense forms, which reveals the bare past suffix *-te*. Because of the zero-exponence, every marker in the segmentation in (2) is already given somewhere in the paradigm in (1) just by the segmentation already present from the word and the stem boundary. Hence, we advance the hypothesis that zero-affixation is a crucial factor that makes subanalysis computationally tractable.

In analogy to the (in-)dependent occurrence of morphemes in syntax, we will call the occurrence of a morpheme as the only form preceding or following the stem *free* and the occurrence as substring of such a string *bound* (hence *-st* is free in (2a) and bound in (2b)). Compare this with the more radical subanalysis in (3) where the apparent 2SG affix *-st* is decomposed into the bound 2SG marker *-s* and the non-first-person affix *-t* which also shows up in the 2PL and the 3SG:

(3) **German verbal agreement with elaborate subanalysis (Müller 2005:10)**

a. PRESENT

| | SG | PL |
|---|---|---|
| 1 | glaub-**e** | glaub-**(e)n** |
| 2 | glaub-**s-t** | glaub-**t** |
| 3 | glaub-**t** | glaub-**(e)n** |

b. PAST

| | SG | PL |
|---|---|---|
| 1 | glaub-**te** | glaub-**te-n** |
| 2 | glaub-**te-s-t** | glaub-**te-t** |
| 3 | glaub-**te** | glaub-**te-n** |

We will argue that it is exactly this boundness contrast that makes the segmentation in (2) (which involves only free forms) uncontroversial and transparent, whereas the further subanalysis in (3) (which adds the bound affix *-s* to the inventory) is more debated and opaque (although, as we will show below, still plausible). In section 3, we will evaluate the effect and search-space-reduction that different degrees of reliance on the free occurrence of a marker have for paradigm learning: If less potential segmentation points are considered, the search space for form-meaning-correspondences remains shallow, and the range of possible subanalyses to be considered is restricted. In fact, each kind of search-space reduction classifies the complexity of subanalyses to be (un-)feasible with a particular strategy.

It is easy to see that the free occurrence of an affix for a specific category $C$ in a given language is closely connected to the fact that other affixal categories $C'$ that

potentially cooccur with $C$ have zero allomorphs. For example, in a language such as German, where verbs are suffixed by tense and agreement affixes, agreement affixes can only be free if there are zero tense affixes (and vice versa). Thus, under our approach zero affixation is a central factor facilitating the learning of subsegmentation. Hence our hypothesis that no language has a subanalysis complexity that demands the learner to explore the full range of possible segmentations has the consequence that zero affixes should be ubiqitous, helping the learner to identify affixes. The final section presents the results of a typological pilot study on the distribution of zero-exponence in tense and agreement across a crosslinguistic language sample, which confirms this prediction.

## 2 Subanalyzing paradigm learning by local optimization

Assuming the learner of a language has already managed to isolate the lexical material of an inflected word form from affixal material by stemming, it is left with what we will call affix strings – everything that precedes and follows the stem (*prefix-string*–stem–*suffix-string*). The verbal agreement paradigm from (3) is strictly suffixing, so our representation can safely omit the empty prefix strings and also the slot-indicating hyphens for the suffixes:[1]

(4) **German verbal agreement suffix strings after stemming**

a.

| PRS | SG | PL |
|---|---|---|
| 1 | e | n |
| 2 | st | t |
| 3 | t | n |

b.

| PST | SG | PL |
|---|---|---|
| 1 | te | ten |
| 2 | test | tet |
| 3 | te | ten |

If subsegmentation is excluded, the meaning assignments to the different markers do not interfere at all. Thus an algorithm without segmentation can simply go through the whole paradigm and identify the most general meaning that exactly matches the distribution of every different affix string. If we prefer general over less general markers (occurring in less cells) and longer over shorter markers (having less segments), the result for (4) is the marker inventory in (5), where preferred form-meaning-correspondences are added before inferior ones.

(5) **Nonsegmenting lexicon assignment preferring perfect generalizations**

(i) *ten*:[–2 +pl +past]
(ii) *te*:[–2 –pl +past]
(iii) *n*:[–2 +pl –past]
(iv) *test*:[+2 –pl +past]
(v) *tet*:[+2 +pl +past]
(vi) *st*:[+2 –pl –past]
(vii) *e*:[+1 –pl –past]
(viii) $t_1$:[+2 +pl –past]
(ix) $t_2$:[+3 –pl –past]

Observe that (5i-iii) cover two paradigm cells, while the other markers do not constitute a generalization as they only repeat the content of a single cell already present in the input. As there is no simple meaning that covers all and only the present tense 2PL and 3SG cells, we also get no generalization for the two cells with the form *t*.

If we allow the form of a marker not only to consist of a full affix strings but also of any of its substrings, different marker hypotheses become mutually exclusive:

[1] As we do not cover morphophonological learning, we assume that the learner is provided with underlying phonological representations.

Based on the established criteria, the possible generalizing marker hypotheses for (4) then are the ones given in (6), again ordered by decreasing quality.

(6) **Possible perfectly accurate generalizations with subanalysis**

(i) *te*:[+past] (iii) *ten*:[–2 +pl +past]
(ii) *n*:[–2 +pl] (iv) *st*:[+2 –pl]

Observe that (6iii) refers solely to full affix strings, and hence already occurrs in (5). The other hypotheses involve the segmentation of at least one cell and are thus excluded without subanalysis. (6i) and (6ii) are clearly compatible with each other: In the two cells where their meanings overlap (past 1PL and 3PL), the affix string *ten* can be consistently segmented into *te* and *n* so that both hypotheses are true. Thus, they can be part of the same analysis. However, they are both incompatible with (6iii), which requires the non-segmentation of exactly these two cells.

In effect, getting a *possible* subanalysis from the set of possible marker hypotheses, involves picking a subset such that *first* all hypotheses are compatible and *second* their combination produces the full content of all paradigm cells. An obvious brute-force strategy for learning would then consist in comparing the set of all possible analyses to each other to identify the globally optimal analysis. However, there is a much more plausible way to find a possible and still quite optimized analysis if we use a greedy strategy – an approach familiar from giving change in everyday life:[2] At each stage, take the set of possible affix hypotheses, choose the optimal one and remove its form from all the cells in the paradigm that are subsumed by its meaning. Recompute the affix hypotheses with the reduced paradigm and repeat the whole procedure until all paradigm cells are empty.[3] If at each step an optimal marker is chosen in a way such that at least one segment is deleted, the algorithm terminates in a finite number of cycles. In this way, the competition between mutually exclusive segmentation options is solved by the optimality criteria for marker hypotheses: Better markers are learned first and competing segmentation options are bled by the removal of forms.

With the optimization criteria established above this spells out as the learning and removal of (6i) and (6ii), which leaves the paradigm from (4) in the following state:

(7) **Reduced affix paradigm after removal of *te*:[+past] and *n*:[–2 +pl]**

a.

| PRS | SG | PL |
|---|---|---|
| 1 | e | |
| 2 | st | t |
| 3 | t | |

b.

| PST | SG | PL |
|---|---|---|
| 1 | | |
| 2 | st | t |
| 3 | | |

[2]Pick the largest denomination not greater than the remaining amount to be made until it is zero.

[3]This type of incremental optimization is closely akin to the Harmonic Serialism version of Optimality Theory (McCarthy 2010) where, in contrast to the standard version of OT, candidates for evaluation may only exhibit a single structural change to the input, but which allows iteration of evaluation cycles, where every cycle takes the output of the preceding cycle as input, up to the point that optimization stagnates, i.e. does not lead to further harmonic improvement. In these terms, the single structural change that defines the optimization cycle for our learner consists of learning a single marker and removing its occurrences from the paradigm.

The next step depends on how exactly we quantify and compare the accuracy and generality measure for marker hypotheses, which we will examine in more detail.

## 2.1 Measuring marker accuracy and generality

To evaluate the accuracy of different form-meaning mappings, we employ standard classification measures used in work on information retrieval and machine learning (Baeza-Yates & Ribeiro-Neto 1999), but it is important to keep in mind that these are simply more formal equivalents of the criteria morphologists use for analysis. (8) exemplifies the use of this terminology with a subset of the possible marker hypotheses for the paradigm in (7), grouped and ordered according to their accuracy and generality.

A *false positive* for an affix hypothesis $H$ of the phonological form $F$ is any paradigm cell for which $H$ is predicted to occur, but which does not contain $F$. Conversely, a *false negative* is a cell where $F$ is predicted *not* to occur by $H$, but still shows up. The meanings of (8a-c) completely accurately characterize the distribution of *st*, *s* and *e* in (7): all cells with the meaning contain the form and conversely all cells containing the form have the meaning (e.g. there is no *st* not covered by [+2 –pl]). Correspondingly, there is an implication both from the meaning to the form ($\leftarrow$) and from the form to the meaning ($\rightarrow$), it is a one-to-one mapping ($\leftrightarrow$).

(8) **Accuracy criteria for the comparison of affix hypotheses**

| | *form*:[meaning] | false positives | false negatives | implica. relation | precision | recall |
|---|---|---|---|---|---|---|
| a. | *st*:[+2 –pl] | – | – | $\leftrightarrow$ | 1 | 1 |
| b. | *s*:[+2 –pl] | – | – | $\leftrightarrow$ | 1 | 1 |
| c. | *e*:[+1 –pl –past] | – | – | $\leftrightarrow$ | 1 | 1 |
| d. | *t*:[+2] | – | yes | $\leftarrow$ | 1 | $\frac{4}{5}$ |
| e. | *t*:[+2 +pl] | – | yes | $\leftarrow$ | 1 | $\frac{2}{5}$ |
| f. | *t*:[–1] | yes | – | $\rightarrow$ | $\frac{5}{8}$ | 1 |
| g. | *t*:[ ] | yes | – | $\rightarrow$ | $\frac{5}{12}$ | 1 |
| h. | *t*:[+1 –pl] | yes | yes | none | $\frac{3}{6}$ | $\frac{3}{5}$ |

With (8d), every cell that matches the meaning [+2] contains the form ($\leftarrow$, perfect precision), yet the occurrence of *t* in the [3 –pl –past] cell is not covered by the marker, resulting in a false negative (false prediction of the non-occurrence of the form) reducing its recall to $\frac{4}{5}$ instead of $\frac{5}{5}$. Whereas (8d-e) do not incur any false positives, (8f-g) do not involve false negatives as they cover all five cells with *t*. On the other hand (8f) leads to false positives for the 3PL and past tense 3SG cells reducing its precision to $\frac{5}{8}$. *Precision* is the fraction of true positives of an affix hypothesis $H$ of form $F$ (the correctly predicted occurrences of $F$) from all paradigm cells matching $H$, and *recall* the fraction of true positives from all occurrences of $F$ in the paradigm. Thus optimizing (and hence maximizing) precision correlates with minimizing false positives, whereas optimizing recall correlates with minimizing false negatives.

It is obvious that these evaluation metrics closely mirror the criteria linguists (and learners of a natural language) employ to determine the correct affix entries for morphological systems. Virtually every morphologist would conclude that (8a) (with perfect precision/recall and zero false positives/negatives) is the correct characterization for the distribution of *st* in (7), and that (8d) (with one false negative and imperfect recall, but perfect precision) is a better analysis for *t* than (8h) (which has two false negatives and also three false positives).[4]

On the other hand, there is no inherent reason in the definition of these criteria to prefer either hypotheses with false positives or false negatives in cases where they can't be completely avoided as is the case with the distribution of *t* in (7). Again this corresponds closely to informed linguistic judgments: Which kind of imperfect distribution is preferable crucially depends on the details of the assumed grammar formalism: The grammar may provide principles or additional machinery that either prevents a marker to occur although it matches a cell (e.g. blocking, impoverishment) or that makes a marker occur in a cell although it does not match its meaning (e.g. rules of referral, empty cells taking the next best marker).

## 2.2 Marker accuracy and generality in incremental learning

We implement the criteria introduced above (where only perfectly accurate marker hypotheses were considered and markers covering more cells and more segments were preferred) with the following optimality-theoretic constraints penalizing false positives and false negatives and honoring marker generality and length:[5]

(9) **Accuracy and generality based constraints selecting affix hypotheses**

| | |
|---|---|
| *OVERINSERTION: | Assign * to every paradigm cell subsumed by the affix hypothesis' meaning not containing its form |
| *UNDERINSERTION: | Assign * to every paradigm cell not subsumed by the affix hypothesis' meaning containing its form |
| MAXCOVERAGE: | Assign ✓ to every paradigm cell subsumed by the affix hypothesis' meaning |
| LENGTH: | Assign ✓ to every segment contained in the form of the affix hypothesis |

(10) **Conservative ranking favoring complete accuracy over generality**

*OVERINSERTION ≫ *UNDERINSERTION ≫ MAXCOVER ≫ LENGTH

With the ranking in (10), (11a) is the optimal hypothesis for the paradigm state in (7): It has neither false positives nor false negatives, covers more cells than (11c) and has more segments than (11b) which has the same meaning and distribution.[6]

[4]Note that these metrics are by themselves completely orthogonal to the issue of deciding on the (non-)segmentation of e.g. *st* as they simply evaluate what is the best characterization for the distribution of a single form and not constitute criteria to compare between different markers.

[5]To allow a more natural formulation, MAXCOVERAGE and LENGTH invoke the assignment of a preference mark ('✓') instead of the standard violation mark ('∗').

[6]Observe that LENGTH ranked lower than the distribution judging constraints has the important effect of preventing the vacuous segmentation of a form into two markers with identical meaning.

(11) **Evaluation of the best affix hypothesis after removal of *te* and *n***

| | *OVERIN | *UNDRIN | MAXCOV | LENGTH |
|---|---|---|---|---|
| ☞ a. *st*:[+2 –pl] | | | ✓$_2$ | ✓$_2$ |
| b. *s*:[+2 –pl] | | | ✓$_2$ | ✓$_1$! |
| c. *e*:[+1–pl–pst] | | | ✓$_1$! | ✓$_1$ |
| d. *t*:[+2] | | $*_1$! | ✓$_4$ | ✓$_1$ |
| e. *t*:[+2 +pl] | | $*_3$! | ✓$_2$ | ✓$_1$ |
| f. *t*:[–1] | $*_3$! | | ✓$_8$ | ✓$_1$ |
| g. *t*:[ ] | $*_7$! | | ✓$_{12}$ | ✓$_1$ |
| h. *t*:[+1 –pl] | $*_3$! | $*_2$ | ✓$_6$ | ✓$_1$ |

Accordingly, the next incremental learning steps consist of adding (11a) and then (11c) to the morpheme inventory, removing them from the paradigm (note that these candidates are compatible, whereas (11a) and (11b) are mutually exclusive). Afterwards *t* is the only form left in the paradigm, hence the segmentation is already fixed. Following our conservative criteria of learning perfect markers wherever possible, we arrive at same segmentation as the minimal subanalysis shown in (2). For our learner, the paradigm state in this fifth cycle repeats the open problem of how to assign meanings in case of an imperfect distribution:

(12) **After removal of *te*:[+past], *n*:[–2 +pl], *st*:[+2 –pl], and *e*:[+1 –pl –past]**

a.

| PRS | SG | PL |
|---|---|---|
| 1 | | |
| 2 | | t |
| 3 | t | |

b.

| PST | SG | PL |
|---|---|---|
| 1 | | |
| 2 | | t |
| 3 | | |

As every possible hypothesis for *t* either has false positives or false negatives, the result clearly depends on wether we judge precision or recall more important than the other. In the optimality-theoretic formalism this is reflected by choosing either the ranking in (13i) or in (13ii):

(13) **Evaluations optimizing (i) precision > recall and (ii) recall > precision**

| (i) | *OVER | *UNDR |
|---|---|---|
| ☞ a. *t*:[+2+pl] | | $*_1$ |
| b. *t*:[–1] | $*_4$! | |
| c. *t*:[+3] | $*_3$! | $*_1$ |

| (ii) | *UNDR | *OVER |
|---|---|---|
| a. *t*:[+2+pl] | $*_1$! | |
| ☞ b. *t*:[–1] | | $*_4$ |
| c. *t*:[+3] | $*_1$! | $*_3$ |

Whether (13a) (no false positives, but one false negative) or (13b) (no false negatives, but false positives) is the better characterization for *t* might be answered differently by morphologists of specific theoretical persuasions. (13a) would a viable analysis for proponents of Paradigm Function Morphology (Stump 2001), which could capture the 'aberrant' occurrence of *t* in the present tense 3SG cell by a rule of referral expanding the distribution of the marker. (13b) would be the option of choice in frameworks which worship underspecification (such as Distributed Morphology, cf. Halle & Marantz (1993); Halle (1997)) and might assume blocking of *t* by more specific *st* and *n* and an Impoverishment rule in the past tense 3SG cell re-

tracting its distribution. To end up with a complete analysis that exactly matches the data, the learner hence has to be adapted to the insertion restrictions and additional mechanisms the grammar employs to cope with false positives or false negatives.

In the following, we simply assume that every marker subsuming a cell's meaning is inserted (i.e. no blocking or other dependencies among inserted markers) which is best matched with a learner that optimizes for precision (prefers markers without false positives as in (13i)). For the conservative ranking in (10) this yields the lexicon in (14), where *t* as a last resort option is separated into two homonyms.

(14) **Segmenting lexicon assignment preferring perfect generalizations**

| | | | | | |
|---|---|---|---|---|---|
| (i) | *te*:[+past] | (iii) | *st*:[+2 –pl] | (v) | $t_1$:[+2 +pl] |
| (ii) | *n*:[–2 +pl] | (iv) | *e*:[+1 –pl –past] | (vi) | $t_2$:[+3 –pl –past] |

While this nicely uncovers the rationale and comparison criteria which underly the minimal segmentation in (2), we also want to follow the path that leads to the more elaborate one from (3) where *st* is subanalyzed into *s* and *t*. The step where this option is abandoned with the current ranking is the evaluation in (11) after the removal of *te* and *n*. In (15), this evaluation is repeated with a slightly different ranking, where the generality of the marker is more important than the avoidance of false negatives (MAXCOVER $\gg$ *UNDERINSERTION). Yet still false positives are avoided wherever possible (undominated *OVERINSERTION).

(15) **Precision and generality biased evaluation after removal of *te* and *n***

| | *OVERIN | MAXCOV | *UNDRIN | LENGTH |
|---|---|---|---|---|
| a. *st*:[+2 –pl] | | ✓$_2$! | | ✓$_2$ |
| b. *s*:[+2 –pl] | | ✓$_2$! | | ✓$_1$ |
| c. *e*:[+1–pl–pst] | | ✓$_1$! | | ✓$_1$ |
| ☞ d. *t*:[+2] | | ✓$_4$ | *$_1$ | ✓$_1$ |
| e. *t*:[+2 +pl] | | ✓$_2$! | *$_3$ | ✓$_1$ |
| f. *t*:[–1] | *$_3$! | ✓$_8$ | | ✓$_1$ |
| g. *t*:[ ] | *$_7$! | ✓$_{12}$ | | ✓$_1$ |
| h. *t*:[+1 –pl] | *$_3$! | ✓$_6$ | *$_2$ | ✓$_1$ |

This metric identifies the largest paradigmatic field, such that all cells contain the same form as being the optimal form meaning mapping (15d). Observe that if the ranking positions of *OVERINSERTION and *UNDERINSERTION are swapped, the *smallest* cell set containing *every* cell of a form is the winner (15f) – which is the metric better matching frameworks with radical underspecification such as DM. With the ranking in (15), the incremental learner finally produces the lexicon in (16) corresponding to the segmentation in (3).

(16) **Segmenting lexicon assignment preferring perfect marker precision**

| | | | | | |
|---|---|---|---|---|---|
| (i) | *te*:[+past] | (iii) | $t_1$:[+2] | (v) | *e*:[+1 –pl –past] |
| (ii) | *n*:[–2 +pl] | (iv) | *s*:[+2 -pl] | (vi) | $t_2$:[+3 +sg –past] |

## 2.3 Algorithm implementation

The pseudo-code in (17) provides the formal framework for out incremental learning algorithm. The effect of different evaluation metrics or purely form-based or meaning-based constraints can be evaluated by parametrization.

(17) **Greedy algorithm for incremental perfect precision learning**

**Input:** a paradigm $P$, i.e. a set of ⟨affix string, meaning⟩-pairs
an empty lexicon $L$

**1** build the set $M$ of all potential markers for $P$

**2** choose the optimal marker $O \in M$ according to the metrics
$\alpha \succ \beta \succ \gamma \succ \gamma$
$\alpha$ minimize the number of false positives
$\beta$ maximize the number of true positives
$\gamma$ minimize the number of false negatives
$\delta$ maximize the number of segments

**3** add $O$ to $L$ **and** remove the affix string of $O$ from all ⟨affix string, meaning⟩-pairs $\in P$ subsumed by its meaning

**4** **if** any ⟨affix string, meaning⟩-pair $\in P$ has a non-empty affix string: **goto** step **1 else** output $L$

There are the following details of the learner which might be fine-tuned: the potential segmentation points assumed when generating possible forms, the possible meanings they are combined with to generate affix hypotheses, and the evaluation metrics for choosing the best among the marker hypotheses.

The consequences of different rankings in the evaluation metrics have been discussed in the previous section. With respect to possible meanings, an aspect not specified explicitly in (17), we remain rather agnostic. However, to learn affix hypotheses which do not incur false positives, the set of meanings is best complete in the sense that it allows to refer individually to any single paradigm cell; in effect there is always a last resort option to build a one-cell marker with perfect precision. Parametrization for possible segmentation points, differences evaluating free vs. bound true positives and false negatives will be introduced in the next section.

# 3 Subanalysis complexity

To identify the globally optimal subanalysis for a paradigm, a naïve algorithm can perform a brute-force search going *firstly* through every possible segmentation and *secondly* through all possible lexicon assignments that match each of the segmentations. For the affix string of a single paradigm cell there are $length(string) - 1$ possible segmentation points. As each segmentation point represents a binary decision of (non-)segmentation, they add up to $2^{length(string)-1}$ possible segmentations. The affix string *test* in (18) for example has $2^{4-1} = 8$ possible segmentations (*test*, *t-est*, *te-st*, *tes-t*, *t-e-st*, *t-es-t*, *te-s-t*, *t-e-s-t*). The number of possible segmentations of the whole paradigm is the product of $2^{(length(string)-1)\cdot number\ of\ occurrences}$ for each affix string: $2_{e}^{0*1} \cdot 2_{st}^{1*1} \cdot 2_{t}^{0*2} \cdot 2_{n}^{0*2} \cdot 2_{te}^{1*2} \cdot 2_{test}^{3*1} \cdot 2_{ten}^{2*2} \cdot 2_{tet}^{2*1} = 2^{12} = 4\,096$. The number of lexicon assignments for a single segmentation is the product of the number of possible partitions for each string: $1_e \cdot 2_{st} \cdot 5_t \cdot 15_n \cdot 203_{te} = 30\,450$ for the seg-

mentation from the solid vertical bars $1_e \cdot 2_s \cdot 52_t \cdot 15_n \cdot 203_{te} = 316\,680$ using all vertical bars in (18).[7]

(18) **Affix string paradigm with two types of segmentation points**

a.

| PRS | SG | PL |
|---|---|---|
| 1 | e | n |
| 2 | s¦t | t |
| 3 | t | n |

b.

| PST | SG | PL |
|---|---|---|
| 1 | te | te¦n |
| 2 | te¦s¦t | te¦t |
| 3 | te | te¦n |

If segmentation is guided by the free occurrences of affix strings, we can distinguish four different types of segmentation points: Full affix strings are already segmented by the word and the stem boundary (19a), so they give possible forms for the last resort of the single non-segmentation of (18). If a segmentation learner only allows these free forms in marker hypotheses (19b), it is restricted to four solid segmentation points in (18) which combine to $2^4 = 16$ possible segmentations. If a learner also allows for the possibility that one part of a subanalysis is a non-affix string (19c), it adds the two interrupted bars in (18) as segmentation points which gives $2^6 = 64$ possible segmentations. Crucially, this adds the possibility of the second order cranberry affix *s* which never occurs without an at least one adjacent other affix. Finally, if we add the remaining six segment transitions as segmentation points (19d), we arrive at the full range of $2^{12} = 4\,096$ segmentations.

(19) **Subanalysis complexity classes as constraints on subaffixes**

a. **Class 0** Affix strings are potential forms (no subaffixes)
b. **Class 1** Every subaffix $S$ of an affix string $AS$ also occurs as an affix string
c. **Class 2** For every binary subanalysis of an affix string $AS$ into $S_1 + S_2$ either $S_1$ or $S_2$ occur as an affix string
d. **Class 3** No restriction on the occurrences of subaffixes

The classes defined in (19) establish an implicational complexity hierarchy regarding the set of possible forms and the set of segmentation points to consider:[8]

(20) **Hierarchy of subanalysis complexity classes**
Class 0 $\subseteq$ Class 1 $\subset$ Class 2 $\subset$ Class 3

The following modifications to the pseudo-code in (17) give different ways to reflect the better accessibility of free affix strings into our incremental learning algorithm.

(21) **Learning algorithm parameters reflecting the free/bound distinction**

a. **step 1:** restrict $M$ to class 0, class 1, or class 2 segmentations of $P$
b. **metric $\beta$:** maximize free true positives $\succ$ bound true positives
c. **metric $\gamma$:** minimize free false negatives $\succ$ bound false negatives[9]

[7] Typically only a subset of these partitions would exclusively consist of sets that correspond to natural classes defined by the meaning system at hand.

[8] Complexity class 2 also restricts possible combinations of segmentation points. Given the free occurrences { *t*, *e*, *st* }, class 2 does not include the segmentations *t-es-t*, *te-s-t* or *t-e-s-t* for *test*.

[9] As the step 1 restriction applies to the current paradigm state, progressing segmentation steps

From the results in the last section we can already conclude that (18) has subanalysis complexity class 1: The non-segmenting analysis (5) fails to identify most of the syncretisms that a class 1 subanalysis finds (14). While a class 2 segmentation adds deeper subanalysis (16) it does not further reduce the number of markers. From a linguistic point of view, we see it as an open question whether the subanalysis in (2) or the one in (3) is more adequate. Still it is a suggestive result that the two major segmentations of German which have been suggested in the literature correspond to the two intermediate degrees of subanalysis complexity we propose. In the next section, we will apply our incremental learner to more complex data and estimate their complexity class.

### 3.1 Subanalysis complexity classes in incremental learning

Swahili verb inflection provides an example, where an adequate subanalysis requires the algorithm to presuppose segmentation points of complexity class 2. Virtually every linguist would concur that the forms in (22) comprise the agreement-tense division we get by parameterizing the algorithm to subanalysis complexity 2: the subjunctive exhibits the bare person/number suffixes, and all other subparadigms combine these with tense/aspect affixes.[10] However, from the perspective of subanalysis complexity, the present and imperfect markers *na* and *li* are cranberry suffixes. Neither of them occurs as a free affix string in any part of the paradigm.

(22) **Swahili verbal agreement (Seidel 1900:10-18), class 2 restricted learner**

| SUB | SG | PL | PRS | SG | PL | IMP | SG | PL |
|---|---|---|---|---|---|---|---|---|
| 1 | ni | tu | 1 | ni-**na** | tu-**na** | 1 | ni-**li** | tu-**li** |
| 2 | u | m | 2 | u-**na** | m-**na** | 2 | u-**li** | m-**li** |
| 3 | a | w-a | 3 | a-**na** | w-a-**na** | 3 | a-**li** | w-a-**li** |

**Lexicon assignment**

(i) *ni*:[+1 –pl] (iv) *li*:[+past] (vii) *m*:[+2 +pl]
(ii) *na*:[–past] (v) *a*:[+3] (viii) *w*:[+3 +pl]
(iii) *tu*:[+1 +pl] (vi) *u*:[+2 –pl]

Consequently, if these data are analyzed with a class 1 restricted learner, it produces the counterintuitive result in (23), which has one marker for every paradigm cell and is thus identical to the unsegmented input paradigm sorted by affix string length.

(23) **Swahili lexicon assignment with class 1 restricted learner**

(i) *nina*:[+1 –pl –pst] (iv) *nili*:[+1 –pl +pst] (vii) *una*:[+2 –pl –pst]
(ii) *tuna*:[+1 +pl –pst] (v) *tuli*:[+1 +pl +pst] (viii) *mna*:[+2 +pl –pst]
(iii) *wana*:[+3 +pl –pst] (vi) *wali*:[+3 +pl +pst] **...** (xviii) *a*:[+3 –pl +sub]

Thus Swahili provides good evidence that class 1 subanalysis is too weak in general, and that at least some languages require subanalysis of complexity 2.

---

may release markers that were bound in previous steps. Note that false positives and true negatives refer to the the non-occurrence of a string, hence we obtain no free vs. bound distinction for them.

[10] Swahili has plenty more verb forms than shown in (22); all of them are transparently structured as the subjunctive and the present. See Seidel (1900) for exhaustive listing.

In fact our impressionistic estimation is that this pattern is even more frequent in other areas of inflection such as adjectival comparison. For example, Persian forms its superlative on top of the comparative:[11]

(24) **Persian adjectival comparison (Mace 2003:53)**

| POSITIVE | COMPARATIVE | SUPERLATIVE | |
|---|---|---|---|
| bozorg | bozorg-**tár** | bozorg-**tar-ín** | 'big' |
| mofid | mofid-**tár** | mofid-**tar-ín** | 'useful' |
| moškel | moškel-**tár** | moškel-**tar-ín** | 'clear' |

Thus the superlative suffix *ín* is again of the cranberry type and effects that the hardly disputable segmentation given in (24) requires class 2 complexity.

Belhare verbal agreement gives some kind of mirror image to Swahili. It does not mark third person agreement but almost entirely has overt tense markers. Thus many of the agreement markers are bound like *ŋa* and *i* in the following subanalysis:

(25) **Belhare intransitive agreement (Bickel 2004:171),[12] class 2 learner**

a.

| PRS | SG | PL |
|---|---|---|
| 1 | t-**ŋa** | t-**i**-**ŋa** |
| 12 | – | t-**i** |
| 2 | (t)-ka | t-**i**-ka |
| 3 | yu | yu |

b.

| PST | SG | PL |
|---|---|---|
| 1 | he-**ŋa** | he-**i**-**ŋa** |
| 12 | – | he-**i** |
| 2 | he-ka | he-**i**-ka |
| 3 | he | he |

**Lexicon assignment**

(i) *he*:[+past]
(ii) *ŋa*:[+1 –2]
(iii) *t*:[–3 –past]
(iv) *i*:[–3 +pl]
(v) *ka*:[–1 +2]
(vi) *yu*:[+3 –past]

Observe that the present tense marker *t* is also bound in this subparadigm. In the course of the incremental learning, the removal of (25i) releases (25ii) which in turn effects the accessibility of *t* as a free form in the third learning cycle.

Finally, the full search space of class 3 is needed to subanalyze a paradigm that always has an overt marker for every category. Recall that our hypothesis is, that such a pattern should not exist. Yet, such complexity can occur if the learner does not 'see' all relevant data i.e. misses the zero-exponent paradigm cells. So if a learner of Swahili has no access to the subjunctive paradigm in (22) but only the present and imperfective, the subanalysis of this highly regular sub-paradigm would be of complexity 3.

The best candidate for data which might only be adequately analyzed by this complexity class is the verb paradigm of the Oceanic language Lenakel (Lynch 1978). Besides additional TAM and subject agreement prefixes (for number) in other positional slots, the core system of verbs is the obligatory combination of an overt agreement prefix with a right adjacent overt TAM prefix as shown in (26). Yet, Lynch (1978:43) states that *ak-* and *ɨm-* "may be omitted in verbs with third person subjects when the context makes the time of action quite clear".

[11] Parallel structures are found e.g. in Ubykh, Sanskrit, Gothic, cf. Bobaljik (2007:12)

[12] The parentheses indicate that *t* surfaces Ø before consonants. Its presence can be attributed to its prevention of intervocalic voicing (e.g. *ti**g**a*), cf. Bickel (2003:551). (25) omits the dual forms.

(26) **Lenakel verbal agreement (Lynch 1978:42-52)**[13]

| | PRES | PAST | STAT | SEQ | NEG |
|---|---|---|---|---|---|
| 1EX | i-ak- | i-~~i~~m- | i-n- | i-ep- | i-~~i~~s- |
| 12 | k-ak- | k-~~i~~m- | k-n- | k-ep- | k-~~i~~s- |
| 2 | n-ak- | n-~~i~~m- | n-n- | n-ep- | n-~~i~~s- |
| 3SG | i-ak- | i-~~i~~m- | i-n- | i-ep- | i-~~i~~s- |
| 3NSG | k-ak- | k-~~i~~m- | k-n- | k-ep- | k-~~i~~s- |
| 3KS | m-ak- | m-~~i~~m | m-n- | m-ep- | m-~~i~~s- |

## 3.2 Zero-marking typology

Up to this point, we have shown that subanalysis complexity has an important potential impact on the performance of learning algorithms for morphological subanalysis. This raises the question whether subanalysis complexity classes have specific empirical consequences. In this section, we report the results of a small typological pilot study on subject agreement and TAM affixes that confirm the linguistic reality of these complexity classes (see Bank & Trommer (to appear) for more details).

Testing the subanalysis complexity of a given inflectional system is a nontrivial task since it would require a complete morphological and phonological analysis of the respective language. Therefore we take occurrence of zero affixes as an indirect indicator for subanalysis complexity: A paradigm which allows a class 2 learner to subanalyze TAM and subject agreement markers with respect to each other must exhibit at least one TAM, or one agreement affix that is zero, whereas a class 1 paradigm must involve at least one zero agreement affix and one zero TAM affix.

Based on Ruhlen's (1987) phyla and macroareas, we have collected verbal paradigms of 20 areally and genetically diverse languages. We have considered only languages with (at least some) subject agreement and TAM inflection on the same side of the stem, disregarding portmanteau expression of subject agreement + TAM, non-finite verb forms, and non-segmental exponence.

Table 1 below shows the results of our survey, where a '+' for Ø-Agr (Ø-TAM) indicates that the language has at least one Ø-affix for subject agreement (TAM); '–' indicates that all relevant affixes of the language are non-zero.

Roughly half of the languages (11/20) have some Ø-marking for subject agreement *and* TAM, and almost all languages (19/20) have some Ø-marking for *either* subject agreement *or* TAM. This result strikingly confirms our predictions.

In fact, since we have not taken into account any other grammatical factors, in the languages under consideration, counting zero morphemes imposes an upper, not a lower bound on the complexity of verbal paradigms. Thus a bound agreement affix (i.e. a marker which only occurs bound to a TAM marker in the verbal paradigm) might also be used as an independent possessive affix or as a free pronoun in the same language. These occurrences which potentially reduce the complexity status of the verbal paradigm would probably be decisive for natural learners of the language, but not show up in our survey. Thus the virtual absence of class 3 languages might actually mean that language learning generally avoids class 3 complexity.

[13] PRES = present, habitual and concurrent mood, STAT = stative/perfective, SEQ = sequential, NEG = negative, KS = known subject; (26) abstracts away from phonological alternations.

**Table 1:** Typological pilot study: language sample

| Language | Phylum | Ø-Agr | Ø-TAM | Source |
|---|---|---|---|---|
| Udmurt | Uralic | + | + | Csúcs (1998) |
| Armenian | Indo-European | + | + | Schmitt (1981) |
| Nahuatl | Uto-Aztecan | + | + | Andrews (1975) |
| Kobon | Trans-N.Gui. | + | + | Davies (1989) |
| Mapudungun | Araucanian | + | + | Zúñiga (2000) |
| Azerbaidjanian | S.Turkic | + | + | Schönig (1998) |
| Turkana | Nilotic | + | + | Dimmendaal (1983) |
| Berber | Afroasiatic | + | + | Kossmann (2007) |
| Choktaw | Muskogean | + | + | Broadwell (2006) |
| Remo | Munda | + | + | Anderson et. al. (2008) |
| Kalkatungu | PamaNyungan | + | + | Blake (1979) |
| Moghol | Mongolian | + | – | Weiers (2011) |
| Belhare | Kiranti | + | – | Bickel (2003) |
| Kannada | S.Dravidian | + | – | Steever (1998) |
| Somali | Cushitic | + | – | El-Solami-Mewis (1987) |
| Inuktitut | Eskimo-Aleut | + | – | Mallon (1991) |
| Swahili | Bantu | – | + | Seidel (1900) |
| Pawnee | Caddoan | – | + | Parks (1976) |
| Manambu | Sepik | – | + | Aikhenvald (2008) |
| Lenakel | CE.M-Polynes. | – | – | Lynch (1978) |

## References


Aikhenvald, Aleksandra Y. 2008. *The Manambu language of East Sepik, Papua New Guinea*. Oxford: Oxford University Press.

Anderson, Gregory D. S., & K. David Harrison. 2008. Remo (Bonda). In *The Munda languages*, ed. by Gregory D. S. Anderson, 557–632. London: Routledge.

Andrews, James Richard. 1975. *Introduction to Classical Nahuatl*. Austin: University of Texas Press.

Baeza-Yates, Ricardo, & Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. New York, NY: ACM Press, Addison-Wesley.

Bank, Sebastian, & Jochen Trommer. to appear. Learning and the complexity of ø-affixation. In *Understanding and measuring morphological complexity*, ed. by Matthew Baerman, Dunstan Brown, & Greville G. Corbett.

Bickel, Balthasar. 2003. Belhare. In *The Sino-Tibetan languages*, ed. by G. Thurgood & R. J. LaPolla, 546–70. London: Routledge.

——. 2004. Hidden syntax in belhare. In *Himalayan languages: past and present*, ed. by A. Saxena, 141–190, Berlin. Mouton de Gruyter.

Blake, Barry J. 1979. *A Kalkatungu grammar*. Canberra: Dept. of Linguistics, Research School of Pacific Studies, Australian National University.

Bobaljik, Jonathan David, 2007. On comparative suppletion. Ms., University of Connecticut.

Broadwell, George Aaron. 2006. *A Choctaw reference grammar*. Lincoln: University of Nebraska Press.

Csúcs, Sándor. 1998. Udmurt. In *The Uralic Languages*, ed. by Daniel Abondolo, 276–304. London and New York: Routledge.

Davies, John. 1989. *Kobon*. London: Routledge.
Dimmendaal, Gerrit Jan. 1983. *The Turkana Language*. Dordrecht: Foris.
El-Solami-Mewis, Catherine. 1987. *Lehrbuch des Somali*. Leipzig: VEB Verlag Enzyklopädie.
Goldsmith, John A. 2010. Segmentation and morphology. In *Handbook of Computational Linguistics and Natural Language Processing*, ed. by Alexander Clark, Chris Fox, & Shalom Lappin. Oxford: Blackwell.
Halle, Morris. 1997. Distributed Morphology: Impoverishment and fission. In *Papers at the Interface*, ed. by Yoonjung Kang Benjamin Bruening & Martha McGinnis, volume 30 of *MIT Working Papers in Linguistics*, 425–449. Cambridge MA: MITWPL.
——, & Alec Marantz. 1993. Distributed Morphology and the pieces of inflection. In *The View from Building 20*, ed. by Kenneth Hale & S. Jay Keyser, 111–176. Cambridge MA: MIT Press.
Harris, Zellig Sabbatai. 1955. From phoneme to morpheme. *Language* 31.190–222.
Kossmann, Maarten G. 2007. Berber morphology. In *Morphologies of Asia and Africa*, ed. by A.S. Kaye, 135–147. Winona Lake IN: Eisenbrauns.
Langer, Hagen, 1991. *Ein automatisches Morphemsegmentierungsverfahren für das Deutsche*. Georg-August-Universität zu Göttingen dissertation.
Lovins, Julie Beth. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics.* 11.22–31.
Lynch, John. 1978. *A grammar of Lenakel*, volume 55 of *Series B*. Pacific Linguistics B.
Mace, John. 2003. *Persian grammar: For reference and revision*. Routledge.
Mallon, Mick. 1991. *Introductory Inuktitut: Reference Grammar*. Montreal: Arctic College - McGill University Inuktitut Text Project.
McCarthy, John, 2010. An introduction to harmonic serialism. Ms., University of Massachusetts, Amherst.
Müller, Gereon, 2005. Subanalyse verbaler Flexionsmarker. Ms. Universität Leipzig.
Parks, Douglas Richard. 1976. *A grammar of Pawnee*. New York: Garland.
Pertsova, Katya. 2011. Grounding systematic syncretism in learning. *Linguistic Inquiry* 42.225–266.
Porter, Martin. 1980. An algorithm for suffix stripping. *Program.* 3.130–137.
Ruhlen, Merritt. 1987. *A Guide to the World's Languages: Classification*, volume 1. Stanford University Press.
Saffran, J. R., E.L. Newport, & R. N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35.606–621.
Schmitt, Rüdiger. 1981. *Grammatik des Klassisch-Armenischen*, volume 32 of *Innsbrucker Beiträge zur Sprachwissenschaft*. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.
Schönig, Claus. 1998. Azerbaidjnian. In *The Turkish Languages*, ed. by Lars Johanson & Éva Ágnes Csató, 248–260. Routledge.
Seidel, August. 1900. *Swahili Konversationsgrammatik*. Heidelberg: Julius Groos.
Steever, Sanford B. 1998. Kannada. In *The Dravidian languages*, ed. by Sanford B. Steever, 129–157. London: Routledge.
Stump, Gregory T. 2001. *Inflectional Morphology*. Cambridge: Cambridge University Press.
Weiers, Michael. 2011. Moghol. In *The Mongolic Languages*, ed. by Juha Janhunen, 248–264. Routledge.
Zúñiga, Fernando. 2000. *Mapudungun*. München: Lincom Europa.