

# On Practical Realisation of Autosegmental Representations in Lexical Transducers of Tonal Bantu Languages

Anssi Yli-Jyrä

Department of Digital Humanities, University of Helsinki  
PL 24, 00014, FINLAND  
anssi.yli-jyra@helsinki.fi

## Abstract

A lexical transducer is a language technology resource that is typically used to predict the orthographic word forms and to model the relation between the lexical and the surface word forms of a morphologically complex language. This paper motivates the construction of tone-enhanced lexical transducers for tonal languages and gives two supporting arguments for the feasibility of finite-state compilation of autosegmental derivations. According to the COMMON TIMELINE ARGUMENT, adding a common timeline to autosegmental representations is crucial for their computational processing. According to the COMPILATION ARGUMENT, the compilation of autosegmental grammars requires combining code-theoretic and model-theoretic research lines.

**Keywords:** lexical transducers, tonal languages, autosegmental phonology, compositionality, origin correspondence

## Résumé (in Swahili)

Makala hii inahamasisha umuhimu wa transdyusa za kimsamiati zinazoainisha pia toni kwa lugha zenye toni, (kama vile lugha nyingi za Kibantu), na inadai kwamba sasa kuna uwezekano wa kuunda fomalismu ya finite-state kwa sarufi za autosegmentali na kwa transdyusa za kimsamiati zinazoainisha pia toni.

## 1. Motivation

*Lexical transducers* (LT) are finite state machines that are specialized to the description of the relation between the word forms and the corresponding morphological analyses. They constitute, for many languages, the main approach to implement the Morphological Component in the conservative BLARK scheme for language technology development. Thus, they are to be considered an essential resource for language technology development of any synthetic (fusional, agglutinative and polysynthetic) language.

In the era of statistical natural language processing and highly successful neural network models, it is relevant to ask whether we still need lexical transducers that are typically constructed manually by a linguist and a lexicographer. The answer is clear: *lexical transducers have particular strengths when there are very limited resources, such as annotated texts, available.* In this context, lexical transducers are very useful as a minimal adequate teacher, or in producing gold morpheme annotation to texts of agglutinative languages in order to train statistical models that learn to produce similar morpheme labels. For the same reason, they have been used to establish an orthography and to maintain the prescriptive spelling of languages that have a scarce written tradition.

Lexical transducers are *more compact, efficient, interpretable, testable and debuggable* than most statistical models. They can be used in checking and validating morphological and phonological grammars and statistical models. Besides this, a lexical transducer can serve as a distilled alternative for a functionally equivalent neural network that has a large memory footprint and whose behaviour needs to be tested and verified against

linguistic parameters and new observations.

*Tone* (Yip, 2002) is a phonologically meaningful pitch distinction that complements vowels and consonants in some spoken languages, forming an autonomous string in the whole phonological representation of sentences. Up to seventy percent of the world's languages are tonal, i.e. they use tone to encode important lexical and grammatical distinctions. To describe the word forms of these languages precisely, one needs to model the relation between the lexical forms and the surface tone melodies. This relation is complex, especially when one considers morphologically rich tone languages in the Bantu family and the context dependent, morpho-syntactic tone alternations.

Because tone disambiguates the meaning between otherwise similar word forms, a *phonologically motivated orthography* explicates the tonal distinctions between similar word forms of a tonal language. This means that if we build a lexical transducer for a tonal language, the computational and linguistic description of its word forms should include the meaningful tone distinctions in the surface word form representation. This is especially important if the morphology is complex and the tone is subject to complex phonological alternations that do not allow treating tone as a segmental feature of the lexical word forms.

A lexical transducer that is describing morphological and morpho-syntactic tone variation will be called *tone-enhanced*. Although tone is not always indicated in the adopted orthographies, tone-enhanced lexical transducers with tone markup are very valuable tools in the development of models that restore the morpho-syntactically determined tone melodies in texts that do not yet indicate tone melody contrasts.

*Construction of a lexical transducer* is typically a very practical effort. Traditionally, practical resource construction has been based on Two-Level Phonology and Morphology, Paradigmatic Morphology, or classical forms of Generative Phonology. More advanced theories such as Autosegmental Phonology, Optimality Theory, Correspondence Theory, Domain Theory, Q-Theory, and Harmonic Serialism do not yet facilitate a regular construction of lexical transducers.

Autosegmental Phonology (AP) (Goldsmith, 1979) is the first major extension of Generative Phonology towards tonal grammars, having such innovations as *morphemic tone* and the *autonomy* of the tonal tier. Today, AP is still one of the most useful and most widely understood phonological theories for the description of tonal alternations in field linguistics. Besides this, AP provides us an important multi-tiered phonological representation that forms a starting point for the development of more recent phonological representations and theories. It is, therefore, natural that we develop some AP-based lexical transducers before trying to construct lexical transducers that are based on more advanced theories.

## 2. The Prior Work

In Computational Morphology and Phonology, one of the most fundamental findings has been that phonological derivations correspond to finite-state relations. This result has led to the development of finite-state methods in natural language processing, including the method that constructs lexical transducers. Muhirwe (2010) treated tone as a segmental phenomenon in a lexical transducer. However, we have not yet been able to compile a large-scale, autosegmental tonal grammar and lexicon into an equivalent finite-state transducer – a tone-enhanced lexical transducer. This is due to three major AP-related computing challenges:

1. the storing of the autosegmental representations
2. the formation of the underlying representations
3. the input-output correspondences in transducers.

Previous research has aimed at *storing* autosegmental representations for algorithmic manipulation. The approaches include codes (Kornai, 1995; Yli-Jyrä, 2015; Jardine and Heinz, 2015; Yli-Jyrä, 2016), strings of tuples (Kiraz, 2000), multi-grained strings (van Leeuwen and te Lindert, 1991; Eisner, 1997; Yli-Jyrä and Niemi, 2006; Barthélemy, 2007; Yli-Jyrä, 2013), tuples of strings (Kay, 1987; Wiebe, 1992), and string sets (Bird and Ellison, 1994). Ideally, the encoding function should be a concatenation homomorphism, but Wiebe (1992) showed that no linear encoding satisfies this requirement when the tiers are not synchronised. However, there are restricted subsets of single-timeline autosegmental representations that are both closed under concatenation and homomorphically i.e. compositionally encodable (Yli-Jyrä, 2015; Jardine and Heinz, 2015; Yli-Jyrä, 2019a).

Computational *formation of underlying representations* (URs) has been addressed in Yli-Jyrä (2013) with a naive and deterministic association rule that works for a finite set of previously known morphemic melodies. Alternatively, the lexicon can consist of sequences of ready-made underlying representations of morphemes or local patterns.

*The input-output correspondences* and the expressive power of autosegmental grammars and their learning problem have been studied recently from the perspective of finite model theory and grammar inference (Jardine, 2014; Jardine, 2017a). Yli-Jyrä (2013) presented a practical transducer compilation method under certain restrictions on tone patterns. This method treated tone and associations naively as span markup in the segmental tier. This paper demonstrates a powerful method for compiling multi-component autosegmental rules over encoded graphs.

## 3. This Paper

Besides giving the motivation to build tone-enhanced lexical transducers, the current short work aims at arguing that the recently found string encoding for arbitrary single-timeline graphs (Yli-Jyrä, 2019b) opens a new technological opportunity. Investment in this opportunity extends the current finite-state technology for lexical transducer construction with the notion of rewritable and constrainable graph structure over a (discrete) timeline.<sup>1</sup>

**THE COMMON-TIMELINE ARGUMENT.** Section 4. argues that (i) by storing the melody and the segments of underlying morphemes on a single timeline, we finally obtain an encoding for the graph structure of the underlying representation, and that (ii) the timeline can be maintained during the phonological processing.

**THE COMPILATION ARGUMENT.** Section 5. argues that (i) a previously developed compilation method for a family of autosegmental rewriting rules can be generalized, but we need (ii) a higher-level logical formalism whose formal semantics links the specified rules into such low-level rules that we currently can compile.

## 4. The Common Timeline Argument

In the construction of lexical transducers, the described sets and relations over phonological representations need to have *good closure properties*. Especially *concatenation* and *Boolean operations* are extremely important.

Untamed autosegmental phonological representations are an impractical idea. They can be viewed naively as a logical description for an equivalence class of graphs or association drawings that differ with respect to the linear representation of the floating tone autosegments and unassociated segments. Such drawings and their classes are deficient wrt closure properties as their concatenation is not compositional (Wiebe, 1992). Moreover, while we can use multi-tape finite-state machines

<sup>1</sup>The current work considers only finite graphs with linearly ordered nodes (=discrete bounded timeline) but is extendible to infinite graphs over a discrete timeline.

to recognize associationless autosegmental representations, such two-tape machines are not closed under Boolean operations as their emptiness is decidable but their equivalence is not (Griffiths, 1968).

Consequently, autosegmental representations need some form of internal *synchronisation* to be computationally well-behaving and closed under important operations. A natural way to introduce synchronisation is to assume morphemic (Yli-Jyrä, 2013) or otherwise local tone association patterns (Jardine and Heinz, 2015; Jardine, 2017b), or by specifying the linearisation of unassociated elements (*inertial* autosegmental representations) (Yli-Jyrä, 2015). With synchronisation, the tiers of the underlying autosegmental representations can be interleaved to a *single timeline*. This common timeline for the underlying tiers should not be confused with the notion of the timing tier. Under a single timeline, autosegmental representations, viewed as graphs, have a string encoding that is bijective and respects concatenation (Yli-Jyrä, 2019b). This enables the description of regular subsets of the code strings and corresponding single-timeline graphs and gives these subsets their good closure properties. The single underlying timeline fixes the origin of tones but allows the *independence* of tiers: floating means lack of association, and linking, shifting, spreading, reduplication and metathesis are processes that alter the association edges without affecting the timeline. The altered associations indicate the remapping of the timeline of the underlying tone tier to the timeline of the segmental tier, but both tiers originate from a shared underlying timeline.

The separation of the origin information from the associations has *advantages*: (i) the shared timeline encodes input-output correspondences. (ii) One may also combine the input and output autosegmental graphs of a phonological mapping and build a union graph that is represented in a single timeline and subject to well-formedness or faithfulness constraints. (iii) One may observe machine learnable local patterns in different stages of the phonological processing, in the origin structure, and in the correspondences.

## 5. The Compilation Argument

Lexical transducers of Bantu languages contain an infinite number of (compound) words and millions of inflections. Due to all these word forms in the transducer, it is not feasible to apply classical graph rewriting methods. Instead, the lexicon has to be constructed by *composition* of regular, possibly infinite relations that are recognized by finite-state transducers. With the single-timeline encoding, we can construct *constraints and conditions of phonological alternations*. In particular, we can

1. compile one-level autosegmental constraints and context conditions, including graph-based local constraints
2. describe feasible changes between input and output autosegmental representations

3. compile constraints on input-output correspondences.

The conditions of rules are the basis for their practical compilation. Yli-Jyrä (2013) presented a method for compiling conditions of autosegmental rewriting rules into finite-state transducers. This method is a combination of (a) optional parallel rewriting with two-level context conditions (aka two-level context restrictions) and (b) a comparison-based output optimisation that converts the optional replacements into obligatory ones.

A tone can have an unbounded number of associations. Multiple association can correspond to edges in a graph (Yli-Jyrä, 2019b) or spans in the timeline (Yli-Jyrä, 2013). Using edges makes the specification of a spreading rule is more complex than the specification where the span boundary is simply moved. Therefore, with the new encoding, a *high-level formalism* for autosegmental rewriting rules is necessary.

It is pretty likely that in real languages, autosegmental representations of their phonology can be embedded to a regular subset of code strings. Regular subsets of this space are closed under all important operations that are needed to define, for these relations, a specification formalism (such as Monadic Second-Order Logic over restricted autosegmental representations) that is mechanically compiled into finite-state transducers. This is the point where the encoding research (Yli-Jyrä, 2019b) can be expected meet research on phonological model-theory and learnability (Jardine, 2017a).

## 6. Conclusion

This article has explained why tone-enhanced lexical transducers are needed and why they can be now be constructed, using a combination of the code-theoretic and model-theoretic perspectives to the AP research. In particular, the article presents two supporting arguments for the feasibility of practical implementation of autosegmental derivations when constructing lexical transducers.

1. The COMMON TIMELINE ARGUMENT of this article states that it is computationally beneficial to assume that there is an underlying timeline that represents the temporal origin of the tones and segments. Having an underlying timeline that is shared through the derivation steps is consistent with the idea of having independent tiers whose associations can be missing and do not need to respect accurately any underlying timeline.
2. The COMPILATION ARGUMENT of this article states that the existing knowledge on rule compilation and the specification formalisms can be applied in order to develop a practical formalism for two-tiered AP.

The author is looking forward to possibilities to implement practical tone-enhanced lexical transducers for tonal languages, especially for morphologically complex ones such as found among the Bantu family that contains some 500 languages.

## 7. Acknowledgements

Writing this article has been enabled by the mobility grant by the University of Helsinki (Faculty of Arts N2/2017), allowing to visit the Hebrew University. The research has had synergy with more general research on string encoding for syntactic and semantic graphs in Research Fellowship projects (279354/273457/313478), and with research on manually aligned parallel texts as a Faculty supported University Researcher (2019). The author started inquiries into the computational implementation of Bantu tone in 2010 under funding of the Academy of Finland, via a Development Research project (2010: 134614 – *MDGs and African language technology: Roadmap to the development of Bantu language resources*). The author is grateful to O. Abend, E. Kuriyozov, J. Tiedemann and C. Gómez Rodríguez and F. Drewes for inspiring discussions, and A. Jardine, A. Fleisch, D. Killian, A. Kornai, and L. Aunio for help with tonal phonology, and A. Hurskainen for the Bantu tone challenge in 2007 and help with the Swahili translation.

## 8. Bibliographical References

- Barthélemy, F. (2007). Multi-grain relations. In *Proceedings of the 12th International Conference on Implementation and Application of Automata*, pages 243–252, Berlin, Heidelberg. Springer-Verlag.
- Bird, S. and Ellison, T. M. (1994). One-level phonology: autosegmental representations and rules as finite automata. *Computational Linguistics*, 20(1):55–90.
- Eisner, J. (1997). Efficient generation in primitive optimality theory. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 313–320, Madrid, Spain, July. Association for Computational Linguistics.
- Goldsmith, J. (1979). The aims of autosegmental phonology. In D. A. Dinnsen, editor, *Current approaches to phonological theory*, chapter 8, pages 202–222. Indiana University Press, Bloomington.
- Griffiths, T. V. (1968). The unsolvability of the equivalence problem for  $\Lambda$ -free nondeterministic generalized machines. *J. ACM*, 15(3):409–413, July.
- Jardine, A. and Heinz, J. (2015). A concatenation operation to derive autosegmental graphs. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, pages 139–151, Chicago, USA, July. Association for Computational Linguistics.
- Jardine, A. (2014). Logic and the generative power of autosegmental phonology. In *Proceedings of the 2013 Annual Meeting of Phonology*.
- Jardine, A. (2017a). The expressivity of autosegmental grammars. Manuscript, June.
- Jardine, A. (2017b). The local nature of tone-association patterns. *Phonology*, 34(2):363–384.
- Kay, M. (1987). Nonconcatenative finite-state morphology. In Bente Maegaard, editor, *3rd Conference of the European Chapter of the Association for Computational Linguistics*, pages 2–10. The Association for Computer Linguistics.
- Kiraz, G. A. (2000). Multitiered nonlinear morphology using multitape finite automata: A case study on Syriac and Arabic. *Computational Linguistics*, 26(1):77–105, March.
- Kornai, A. (1995). *Formal Phonology*. Garland Publishing, New York.
- Muhirwe, J. (2010). Morphological analysis of tone marked Kinyarwanda text. In *Finite-State Methods and Natural Language Processing*, volume 6062 of *Lecture Notes in Computer Science*, pages 48–55. Springer Berlin Heidelberg.
- van Leeuwen, H. and te Lindert, E. (1991). Speech maker: text-to-speech synthesis based on a multi-level, synchronized data structure. In *International Conference on Acoustics, Speech, and Signal Processing, 1991*, pages 781–784 vol.2, Apr.
- Wiebe, B. (1992). Modelling autosegmental phonology with multitape finite state transducers. Master’s thesis, Simon Fraser University.
- Yip, M. (2002). *Tone*. Cambridge Studies in Linguistics. Cambridge University Press.
- Yli-Jyrä, A. and Niemi, J. (2006). Pivotal synchronization languages: A framework for alignments. In Anssi Yli-Jyrä, et al., editors, *Finite-State Methods and Natural Language Processing*, volume 4002 of *Lecture Notes in Computer Science*, pages 271–282. Springer Berlin Heidelberg.
- Yli-Jyrä, A. (2013). On finite-state tonology with autosegmental representations. In Mark-Jan Nederhof, editor, *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, St. Andrews, Scotland, UK, July 15-17, 2013*, pages 90–98. The Association for Computer Linguistics.
- Yli-Jyrä, A. (2015). Three equivalent codes for autosegmental representations. In Thomas Hanneforth et al., editors, *Proceedings of the 12th International Conference on Finite-State Methods and Natural Language Processing, Düsseldorf, Germany, June 22-24, 2015*. The Association for Computer Linguistics.
- Yli-Jyrä, A. (2016). Aligned multistring languages. In *TTATT 2016, Proceedings of the Workshop, Workshop on Trends in Tree Automata and Tree Transducers*, Seoul, South Korea, July 18.
- Yli-Jyrä, A. (2019a). *Optimal Kornai-Karttunen Codes for Restricted Autosegmental Representations*. Number 224 in CSLI Lecture Notes. CSLI Publications, Stanford, USA.
- Yli-Jyrä, A. (2019b). Transition-based coding and formal language theory for ordered digraphs. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 118–131, Dresden, Germany, September. Association for Computational Linguistics.