Topology of Language Classes

Sean A. Fulop
Dept. of Linguistics
California State University Fresno
sfulop@csufresno.edu

David Kephart
Link-Systems International
dkephart@link-systems.com

Abstract

The implications of a specific pseudometric on the collection of languages over a finite alphabet are explored. In distinction from an approach in (Calude et al., 2009) that relates to collections of infinite or bi-infinite sequences, the present work is based on an adaptation of the "Besicovitch" pseudometric introduced by Besicovitch (1932) and elaborated in (Cattaneo et al., 1997) in the context of cellular automata. Using this pseudometric to form a metric quotient space, we study its properties and draw conclusions about the location of certain well-understood families of languages in the language space. We find that topologies, both on the space of formal languages itself and upon quotient spaces derived from pseudometrics on the language space, may offer insights into the relationships, and in particular the distance, between languages over a common alphabet.

1 Introduction

The question of distance between languages, and comparison of possible definitions, has relatively less consideration in the literature than other language issues, with notable exceptions being (Berstel, 1973) and (Salomaa and Soittola, 1978). This may seem surprising, considering that the current digital climate necessitates the measurement of likeness between texts and languages, for instance in search engine entries and results. *Ad hoc* measures of differences exist based upon rooted tree distances, but these are more like attempts to incorporate the intuitive notion of differences between words than overall differences between languages. In Linguistics, as well, there is as yet no accepted way of measuring the distance between two dialects of a language, with each employing the same vocabulary.

This paper borrows a pseudometric from cellular automata theory to use language density and form a topology on the set of languages (consisting of words of finite length over a fixed alphabet). A similar pseudometric is discussed in (Cattaneo et al., 1997). Our goal is to continue a systematic review and categorization of language distances, with a view to determining what gives rise to apparent weaknesses and strengths of each. As seen in (Salomaa and Soittola, 1978) and (Yu, 1997) language *density* is understood as the number of words in a language, conceived of as a function $\varrho(n)$ of word length n. This is shown to convey information about the nature of a language. Analysis of language density over finite words may be confined to the treatment of regular languages (Yu, 1997; Kozik, 2005), or seen as a probability density of distance between infinite sequences (Kozik, 2006).

Herein we continue the approach of (Kozik, 2006) of capturing distances between arbitrary languages, specifically by looking at features of the topology generated by each. We consider that languages—natural or formal—are most beneficially understood as potential or actually infinite objects. As such, language patterns may or may not be adequately defined syntactically. We continue the work of Kozik in grasping language differences as word density of distinctions at the limit. Since such a limit may not exist, we look at a pseudometric inspired by Besicovitch (1932) that captures, in fact, the upper density limit of language differences.

Next, we consider "where," in the resulting "Besicovitch" topology of the language space, individual

Proceedings of the 14th Meeting on the Mathematics of Language (MoL 14), pages 26–38,
Chicago, USA, July 25–26, 2015.

languages lie. We also look at how this relates to the Chomsky hierarchy of languages. We find that the pseudometric space of languages is not complete, and look at the lifting of the pseudometric to a quotient metric space. The hope is that this consideration may contribute to a list of relative advantages and disadvantages associated with various candidate language topologies. Our contribution is thus conceived as a part of a broader exploration in search of the most useful topology of language spaces, with eventual application to linguistic problems like measuring the distance between dialects over a common vocabulary. We have tried to study the Besicovitch topology and its quotients in some detail, but some proofs have been condensed to outlines due to space limitations.

1.1 Early approaches

Nelson (Nelson, 1980) elaborated work by Walter (Walter, 1975) which constructed a topological space from a space of rewriting grammars by means of successive divisors of grammatical derivations. The resulting topologies of both languages and grammatical derivations are equivalent to quasi-ordered sets, and have the property that each point has a smallest open neighborhood. If such a topology is T_1 then it is discrete.

An equivalence relation between languages was suggested by Marcus (Marcus, 1966; Marcus, 1967) based on equivalence of word contexts. Improved and elaborated by Dincă (Dincă, 1976), this approach treats the space of languages as a semigroup over the alphabet, and a distance in the quotient space (dividing by context equivalence) measures the distance between context classes of strings with respect to some chosen language.

The above described approaches, while not without interest where linguistic applications are in view, do not yield a "sufficiently smooth" topology of a language space. The first approach similar in spirit to our main thread was published by Vianu (1977), who applied the metric proposed earlier by Bodnarchŭk (1965). This approach has a number of variants, but we will point out the most important conclusions to be drawn from them as well as possible limitations of this approach.

1.2 Current literature on language topologies and distances

Language spaces allowing infinitary words, on the other hand, can be more easily endowed with adequate topologies arising out of the word topology (Calude et al., 2009), but this will not be a topic of discussion here because there seems to be no application of infinitary word languages to the study of natural human languages.

2 Preliminaries

In this section we review some basic definitions from formal language theory and review the best-known approach to language distance, namely, what we will call the Cantor metric.

2.1 Notation and Definitions

For the most part, we adopt notation common to formal language theory. There are a few modifications in the interest of brevity and, hopefully, clarity of expression. We consider a language as a set of *words* which are concatenated from symbols in an *alphabet* Σ with finite cardinality α. We will deal only with words of finite length (as opposed to the words discussed, for instance, in (Calude et al., 2009)). By a *language space* we mean the collection of all possible languages, namely, 2^{Σ^*}.

Sets. We frequently employ *the symmetric set difference of sets A and B*, denoted $A \triangle B$.

Words. The length of word w will be denoted $|w|$ and will always be non-negative. The empty word, which is the unique word of length zero, will, as usual, be denoted λ. When we need to refer to the ith symbol of the word w, we will denote this by $w_{[i]}$, preserving ordinary subscripts for the enumeration of words. The fundamental operation on symbols is (non-commutative) concatenation, which is represented multiplicatively. We use the Kleene-$*$ (-star) and -$+$ (-plus) operations in the usual way. Moreover, Σ^n denotes the set of all words of length n, and $\Sigma^{<n}$ denotes the set of all words of lengths up to $n-1$.

Languages. The empty language is simply the null set, $\emptyset$. Concatenation extends from words to languages. That is, if L and M are languages, then $LM = \{uv : u \in L, v \in M\}$. Suppose L is a language over Σ and $n \in \mathbb{N} \cup \{0\}$. Then we de-

note by L^n (respectively $L^{<n}$, for $n > 0$) the set $L \cap \Sigma^n$ (respectively, $\bigcup_{i=0}^{n-1} L^i$). For example, L^0 is either $\emptyset$ or $\{\lambda\}$. The *density of language* L is the sequence $\{\varrho_n\}_{n\in\mathbb{N}}$ such that $\varrho_i = |L^i|$. Then $|L^{<n}|$ is the nth partial sum of the series $\sum \varrho$. Finally, given languages L and M, we denote by $L \triangle^n M$ (respectively, $L \triangle^{<n} M$) the symmetric set difference between words of length n (respectively, less than n) in the two languages.

Remark 1. *Note that*

$$|\Sigma^{<n}| = \frac{\alpha^n - 1}{\alpha - 1}. \tag{1}$$

2.2 Language norms, metrics and the Cantor space

In setting out to find ways to adequately express the "distance" between two languages, we consider how to adapt the notions of size and separation into the realm of formal symbols. We already observe that the first defined language distance, i.e., the distance between two languages, in the literature, derives from the density of their symmetric set difference. The metric mentioned by (Vianu, 1977) is based on the shortest word in $L \triangle M$. Indeed, this leads to a full metric, and a metric topology on 2^{Σ^*}. By analogy to the norm in a normed space representing distance from a zero point, and hence magnitude, a "language norm" can be elaborated from a pseudometric.

The reader will recall that a pseudometric d on space X is a function that maps $X \times X$ to $\mathbb{R}^{\geq 0}$, such that $d(x,x) = 0$, $d(x,y) = d(y,x)$ and $d(x,y) + d(y,z) \geq d(x,z)$. We call a pseudometric a *language distance* just in case it additionally is such that, if $L \cap N = \emptyset$ and $M \subseteq N$, then $d(L,M) \leq d(L,N)$.

Then, to every language distance we may associate a function $\|\cdot\|_d : 2^{\Sigma^*} \to \mathbb{R}^{\geq 0}$ by defining $\|L\|_d = d(L,\emptyset)$. Note that $\|\cdot\|_d$ has the following properties:

$$\|\emptyset\|_d = 0 \tag{2}$$

$$\|L \cup M\|_d \leq \|L\|_d + \|M\|_d \tag{3}$$

$$L \subseteq M \implies \|L\| \leq \|M\| \tag{4}$$

We define a *language norm* as any such function on languages.

Lemma 2. *To each language norm $\|\cdot\|$ there corresponds a unique language distance d such that $d(L,M) = \|L \triangle M\|$.*

The contrapositive of Lemma 2 also holds. That is, for any language distance d on 2^{Σ^*}, the function $\|\cdot\| : 2^{\Sigma^*} \to \mathbb{R}^{\geq 0}$ such that $\|L\| = d(L,\emptyset)$ defines a unique language norm.

2.3 The Cantor language metric and topology on 2^{Σ^*}

A Cantor language space

Two languages can be compared by beginning with the shortest word in each language and proceeding to longer words. A first notion of distance is obtained using the word-length of the first distinction between languages so observed. To this end, let the language space then be normed by assigning the norm 0 to $\emptyset$ and by associating each non-empty language to a power of $1/2$, as follows.

Definition 3. *The language norm $\|\cdot\|_1 : 2^{\Sigma^*} \to \mathbb{R}$ is as follows: for $L \in 2^{\Sigma^*}$,*

$$\|L\|_1 = \begin{cases} 0 & \text{if } L = \emptyset, \\ 2^{-\min\{|w| : w \in L\}} & \text{otherwise.} \end{cases} \tag{5}$$

Observe that $-\log_2 \|L\|_1 \in \mathbb{N}$ for all non-empty L.

To this language norm corresponds the following language metric.

Definition 4. *The function $d_1 : 2^{\Sigma^*} \times 2^{\Sigma^*} \to \mathbb{R}$ is a metric, where, for L and M in 2^{Σ^*}, $d_1(L,M) = \|L \triangle M\|_1$*

To see that d_1 is in fact not only a pseudometric but a metric, consider that $d_1(L,M) = 0$ iff $L \triangle M = \emptyset$, i.e., iff $L = M$. Let τ_1 be the metric topology induced on 2^{Σ^*} by d_1. For reasons to be made clear below, we call $\|\cdot\|_1, d_1$, and τ_1 the *Cantor norm, distance,* and *topology,* respectively, on a language space.

The open neighborhoods of radius $\epsilon > 0$ around some language $L \in 2^{\Sigma^*}$, denoted $\mathcal{B}_\epsilon(L) = \{M \in 2^{\Sigma^*} : d_1(L,M) < \epsilon\}$, form the standard basis for τ_1. Since distances between distinct languages are powers of $1/2$, it follows that elements of the standard metric basis for $(2^{\Sigma^*}, \tau_1)$ form the collection

$$\mathcal{C} = \{\mathcal{B}_{2^{-n}}(L) : n \in \mathbb{N}, L \in 2^{\Sigma^*}\}. \tag{6}$$

Definition 5 ((Vianu, 1977; Genova and Jonoska, 2006)). *The* language cylinder set $C_{L,k}$ *of length* $k \in \mathbb{N}$ *around language* $L \in 2^{\Sigma^*}$ *is:*

$$C_{L,k} \stackrel{\text{def}}{=} \{M \in 2^{\Sigma^*} : L \cap \Sigma^{\leq k} = M \cap \Sigma^{\leq k}\}. \quad (7)$$

Now let $\mathcal{C}_k \stackrel{\text{def}}{=} \{C_{L,k} : L \in 2^{\Sigma^*}\}$ be the collection of all language cylinder sets of length k. From (6) and (7) it follows that the collection $\mathcal{C} \stackrel{\text{def}}{=} \bigcup_{k\in\mathbb{N}} \mathcal{C}_k$, comprising all language cylinder sets, is the standard basis for $(2^{\Sigma^*}, \tau_1)$.

Cantor topology on a language space

As it turns out, τ_1 is equivalent to the topology of the Bodnarchŭk metric space discussed in (Vianu, 1977). We quickly recap the properties of this topology on a language space, as proven in (Vianu, 1977) and (Genova and Jonoska, 2006).

Lemma 6. *In* $(2^{\Sigma^*}, \tau_1)$*, every cylinder set is both closed and open.*

Lemma 7. *A sequence* $\{L_i\}_{i\in\mathbb{N}} \subset 2^{\Sigma^*}$ *converges to language* L *in* $(2^{\Sigma^*}, \tau_1)$ *iff for all* $m \in \mathbb{N}$, $|(L_i \triangle^m L)| = 0$ *for all but finitely many* i*. In this case we write* $L_i \to L$*.*

From this and the fact $L \cap \Sigma^{\leq i} \to L$ we also have:

Corollary 8. *The finite languages are dense in a space of languages under the* τ_1 *topology.*

Lemma 9. *The topological space* $(2^{\Sigma^*}, \tau_1)$ *is homeomorphic to the Cantor space.*

Thus the terminology "Cantor language space, topology," etc.[1]

Corollary 10. $(2^{\Sigma^*}, \tau_1)$ *is compact, perfect, and totally disconnected.*

3 Besicovitch pseudometric, language norm and topology

We now consider a language distance that is in many respects more satisfactory than the Cantor distance d_1, by exploiting the general philosophy of comparing languages by comparing finite sections of languages. We then show several results, including that neither finite nor locally testable languages are dense in the topology induced. We call this alternative pseudometric the Besicovitch distance, denoted by d_ζ. Under the topology induced, a language space is not compact. Rather, it has a geometry which becomes apparent from the vantage point of a metric quotient space.

The original Besicovitch pseudometric expressed the distance between two almost-periodic real-valued functions (Besicovitch, 1932) $\phi, \psi \in \ell^1$ as

$$d_{B^p}(\phi, \psi) \stackrel{\text{def}}{=} \limsup_{n\to\infty} \frac{1}{2n+1} \sum_{-n}^{n} |\phi(x) - \psi(x)|.$$

Because this pseudometric depends on the evaluation of the two functions only at discrete intervals, it is naturally adaptable to expressing distances between objects with a bound proportion of differences, as with the distance between cellular automata (Cattaneo et al., 1997); our adaptation to languages is in some sense a generalization thereof.

3.1 A Besicovitch pseudometric on language spaces

We begin by defining a Besicovitch-style language norm. Rather than halting at a particular term of the density of the symmetric set difference between two languages, this norm considers the derived infinite series $|L \triangle^{<n} M|$ in ratio to the total possible words over Σ^* (given by (1)) as n goes to infinity.

Definition 11. *Let* $\|\cdot\|_\zeta$ *for fixed alphabet* Σ *be the function defined:*

$$\|L\|_\zeta \stackrel{\text{def}}{=} \limsup_{k\to\infty} \frac{|(L \cap \Sigma^{<k})|}{|\Sigma^{<k}|} \quad (8)$$

We call $\|\cdot\|_\zeta$ *the* Besicovitch language norm*.*

Then let d_ζ *be the function mapping* (L, M) *to* $\|L \triangle M\|_\zeta$*, and call it the* Besicovitch language distance*.*

By Lemma 2, distance d_ζ is a language pseudometric.

Remark 12. *The Besicovitch distance* d_ζ *between languages*

[1] As pointed out by an anonymous reviewer, the Cantor topology can be understood as the profinite completion of an algebra of recognizable languages (Gehrke, 2009). While this does not modify the topological characteristics of the space, it does raise the interesting point that the Besicovitch topology, the main subject of this paper, likely cannot be so conceived: the Besicovitch metric (quotient) space is not complete, by Corollary 31.

1. *can be described as the upper density of their set-difference;*

2. *turns out to constitute (like the Besicovitch language norm) a continuous, surjective mapping of 2^{Σ^*} into the unit interval $[0,1]$;*

3. *for a language and its complement is 1, since $(L \cap \Sigma^k) \triangle (\neg L \cap \Sigma^k) = \Sigma^k$ for every $k \in \mathbb{N}$;*

4. *constitutes a strict pseudo-metric if $|\Sigma| > 1$, since, for instance, $\|M\|_\zeta = 0$ where $M = \{a\}$, so $d_\zeta(M, \emptyset) = 0$ even though $M \neq \emptyset$;*

5. *given languages $L, M \in 2^{\Sigma^*}$, can be written as follows:*

$$d_\zeta(L, M) = \limsup_{k \to \infty} \left| L \triangle^{<k} M \right| \left(\frac{\alpha - 1}{\alpha^k - 1} \right)$$

We present the following without proof.

Lemma 13. *If L and M are disjoint languages in 2^{Σ^*}, then $\|L\|_\zeta + \|M\|_\zeta = \|L \cup M\|_\zeta$.*

Corollary 14. *For $L, M \in 2^{\Sigma^*}$, $\|L\|_\zeta + \|M\|_\zeta = \|L \cup M\|_\zeta$ if and only if $\|L \cap M\|_\zeta = 0$.*

Corollary 15. *For all $L, M \in 2^{\Sigma^*}$, $\|L\|_\zeta + \|M\|_\zeta \geq \|L \cup M\|_\zeta$.*

The conclusion here is that $\|\cdot\|_\zeta$ is truly "norm-like." For the remainder of this section, we drop most subscripts ζ.

To establish surjectivity, we first need a way to construct a language with a specified arbitrary norm.

Definition 16. *Given $0 \leq r \leq 1$, consider the sequence $\check{r}_{\langle\alpha\rangle} \stackrel{\text{def}}{=} \{\lfloor r\alpha^k \rfloor\}_{k \in \mathbb{N}}$. Then we call $\mathbf{L}_r$ the set of r-*simple languages *in 2^{Σ^*}, defined as follows:*

$$\mathbf{L}_r \stackrel{\text{def}}{=} \left\{ L \in 2^{\Sigma^*} : \{|(L \cap \Sigma^i)|\}_{i \in \mathbb{N}} = \check{r}_{\langle\alpha\rangle} \right\}.$$

Lemma 17. *For $r \in [0,1]$ there is at least one r-simple language; moreover for each particular value r, every r-simple language has norm r.*

Proof. By construction, for each $r \in [0,1]$ the sequence $\check{r}_{\langle\alpha\rangle}$ exists. We can select $\check{r}_k$ words in Σ^k for all k. This amounts to the construction of a language L in $\mathbf{L}_r$ for each $r \in [0,1]$. But then $\|L\| = r$, which establishes the claim. □

Now we have established our hoped-for result.

Corollary 18. *The Besicovitch language norm is a surjective mapping from 2^{Σ^*} onto $[0,1]$.*

In addition, it is relatively easy to see that diagonalization yields that there are uncountably many r-simple languages for each $r \in [0,1]$.

3.2 Besicovitch distance quotient space

The Besicovitch distance equivalence induces a quotient space on 2^{Σ^*}

We next form collections of languages at distance zero from each other and map each such collection to a point in a quotient space, which can then be metrized. So, given $L, M \in 2^{\Sigma^*}$, let $L \sim_\zeta M$ if $d_\zeta(L, M) = 0$.

Proposition 19. *The relation $\sim_\zeta$ is an equivalence on 2^{Σ^*}.*

Proof. Reflexivity and symmetry are apparent and, if $L, M, N \in 2^{\Sigma^*}$ such that $L \sim M$ and $M \sim N$, then $0 = d(L, M) + d(M, N) \geq d(L, N)$. From Remark 12(1), $L \sim N$. □

The collection of $\sim$ equivalence classes will be called the *Besicovitch quotient space* over 2^{Σ^*}, denoted $\mathfrak{Q}_\zeta^\Sigma$. Here we will drop the ζ subscript for notational clarity and assume the language space 2^{Σ^*} unless otherwise noted. Elements of the quotient space (points in $\mathfrak{Q}$) will be denoted with sans-serif letters $\mathsf{L}, \mathsf{M}, \mathsf{N}, \ldots$, while collections of such points will be denoted with corresponding bold letters $\mathbf{L}, \mathbf{M}, \mathbf{N}, \ldots$. Let $\eta \colon 2^{\Sigma^*} \to \mathfrak{Q}_\zeta^\Sigma$ denote the quotient mapping which takes a language to its $\sim$ equivalence class.

Remark 20. *As a partition of 2^{Σ^*}, the mapping η is well-defined and surjective, but not injective since it is a quotient mapping. The set operations of union, intersection and complementation are preserved by mappings from collections of points in $\mathfrak{Q}$ to the sets of languages of which they are equivalence classes. In particular, every topology on $\mathfrak{Q}_\zeta^\Sigma$ is the quotient of a topology on 2^{Σ^*}.*

When language L is a member of language family $\mathbf{L}$, and every member of $\mathbf{L}$ is contained in an equivalence class in the collection of points $\mathbf{L} \subseteq \mathfrak{Q}_\zeta$, we will write $L \in \mathsf{L}$ and $\mathbf{L} \subseteq \mathbf{L}$ instead of the more tedious $\eta(L) = \mathsf{L}$ and $\eta(\mathbf{L}) \subseteq \mathbf{L}$.

Lemma 21. *For languages $L, M \in 2^{\Sigma^*}$, $L \nsim M$ iff there exists a positive integer m such that, for infinitely many word-lengths n, $|(L \triangle^{<n} M)| \geq |\Sigma^{<n-m}|$.*

Proof. ($\Rightarrow$) Suppose there is no such m. That would mean that, for each $m \in \mathbb{N}$, there is a word length n_m such that $k > n_m$ implies $|(L \triangle^{<k} M)| < |\Sigma^{k-m}|$. We can then construct an increasing sequence $\{k_i\}_{i\in\mathbb{N}}$ where $k_0 = 0$ and k_i ($i > 0$) is the least integer greater than k_{i-1} such that $|(L \triangle^{<k'} M| > |\Sigma^{<k'-i}| = \sum_{j=0}^{k'-i-1}|\Sigma^j| = \frac{\alpha^{k'-i}-1}{\alpha-1}$ if $k' > k_i$. But, if this were true, then the Besicovitch distance between the two languages would be 0, since a straightforward calculation shows that, for each $m \in \mathbb{N}$, $d_\zeta(L, M)$ is bounded above by α^{-m}.

($\Leftarrow$) Assume that, for some $m \in \mathbb{N}$ and for n sufficiently large, $|(L \triangle^{<n} M)| \geq |\Sigma^{n-m}|$. Then a similarly straightforward calculation shows that $\|L \triangle M\| = \limsup_{k\to\infty} \frac{|L\triangle^{<k}M|}{|\Sigma^{<k}|}$ is bounded below by, for instance, $\alpha^{-m}/2$. Thus, $L \nsim_\zeta M$.

Note that when two languages are similar, the sequence used in the first part of the proof is *finite*. □

Definition 22. *Given languages $L, M \in 2^{\Sigma^*}$, we will denote by $K_\zeta(L, M)$ the (possibly finite) increasing integer sequence $\{k_i\}_{i\in\mathbb{N}}$ in accord with the above lemma. Indeed, $K_\zeta(L, M)$ is infinite precisely when $L \sim M$.*

We note that if $K_\zeta(L, M)$ has at least i terms, then $k_i > i$ and, by considering the words in $L \triangle M$ of length greater than m_i, we have a first estimate of the distance between two languages, namely, α^{-i}.

By Lemma 21, the unique sequence $K_\zeta(L, M)$ expresses the relative location of languages in the quotient space $\mathcal{Q}_\zeta$.

The quotient space has a natural metric quotient topology

We define the metric $\mathbf{d}_\zeta$ on the Besicovitch quotient space as the lifting of Besicovitch distance d.

Definition 23. *Let the distance $\mathbf{d}_\zeta$ between points L and M in $\mathcal{Q}_\zeta$ be set equal to* $\inf\{d(L, M)\colon L \in \eta^{-1}(\mathsf{L}), M \in \eta^{-1}(\mathsf{M})\}$.

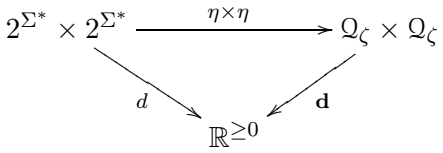
Lemma 24. *For $\mathsf{L}, \mathsf{M} \in \mathcal{Q}_\zeta$, $\mathbf{d}_\zeta(\mathsf{L}, \mathsf{M}) = 0$ iff $\mathsf{L} = \mathsf{M}$.*

Proof. Only the implication left to right requires a proof. Suppose that $\mathbf{d}(\mathsf{L}, \mathsf{M}) = 0$. Now suppose, contrary to the claim, that there is some language $L \in \mathsf{L}$ but not M. We conclude from the preceding definition 23 that there exists $M \in \mathsf{M}$ such that $d(L, M) = \epsilon > 0$. But then, for arbitrary languages $L' \in \mathsf{L}$ and $M' \in \mathsf{M}$, the triangle inequality provides us that

$$\begin{aligned} \epsilon \leq d(L, M) &\leq d(L, L') + d(L', M') + d(M', M) \\ &\leq d(L', M'). \end{aligned} \tag{9}$$

Thus $\mathbf{d}(\mathsf{L}, \mathsf{M}) \geq \epsilon/2 > 0$ by the preceding definition, Q.E.A. □

Corollary 25. *For $L, M \in 2^{\Sigma^*}$ the diagram below commutes, showing the isometry between Besicovitch language space and quotient space.*



It is by now evident that the Besicovitch quotient space is a metric space under distance $\mathbf{d}$. Moreover we have:

Corollary 26. *If languages L, M are in point $\mathsf{L} \in \mathcal{Q}_\zeta$, then $\|L\| = \|M\|$.*

This just means that the $\mathbf{d}$ metric topology on $\mathcal{Q}_\zeta$ is the quotient of the pseudo-metric topology induced by d on 2^{Σ^*}. Let $\tilde{\tau}_\zeta$ denote the collection of open sets in $\mathcal{Q}_\zeta$ under the $\mathbf{d}$ metric topology, and let τ_ζ denote the collection of language sets in 2^{Σ^*} such that $\eta(\tau_\zeta) = \tilde{\tau}_\zeta$. We will call τ_ζ the *Besicovitch language topology.*

Convergence has a novel interpretation in the quotient space

From Remark 12, the Besicovitch language topology is not T_1, and so convergence to a language is not well-defined in $(2^{\Sigma^*}, \tau_\zeta)$. But there is no such difficulty with the quotient space.

Lemma 27. *A sequence $\{\mathsf{L}_i\}_{i\in\mathbb{N}}$ in $\mathcal{Q}_\zeta$ converges to the point $\mathsf{L} \in \mathcal{Q}_\zeta$ iff the following: $\forall m \in \mathbb{N}\ \exists k_m \in \mathbb{N}$ such that $i > k_m$ means that, if language $L_i \in \mathsf{L}_i$ and $L \in \mathsf{L}$ then there exists integer N_i for which $k > N_i$ implies $|(L \triangle^k L_i)| < \alpha^{k-m}$.*

Note that, unlike the case of the Cantor space, in the Besicovitch language (quotient) topology, a convergent sequence of points converges to $\sim$ equivalence with the (languages in the) limit point.

The quotient space is perfect but not compact

We can next address the compactness question for the Besicovitch quotient space, since it is a metric space, by determining whether every infinite sequence of points has a convergent subsequence. We ultimately show here that neither $\mathcal{Q}_\zeta$ nor 2^{Σ^*} is compact, although $\mathcal{Q}_\zeta$ is a perfect set.

We first establish the latter property using the following fact.

Lemma 28. *Every point in $(\mathcal{Q}_\zeta, \tilde{\tau})$ is a condensation point.*

Proof. Every open set M in $(\mathcal{Q}_\zeta, \tilde{\tau})$ includes the image of an open interval in the subset topology of $[0, 1]$ and, by a diagonalization, is uncountable. Each number in M has a distinct open inverse image in $(\mathcal{Q}_\zeta, \tilde{\tau})$. □

It then follows immediately that $\mathcal{Q}_\zeta$ is perfect. To make progress on the compactness question, we construct a family of two-sided *word ideals* in Σ^* which, when split into non-disjoint right ideals, yields an infinite sequence in the quotient space with no convergent subsequence. We will call I a right, left, or two-sided *word ideal* of the monoid Σ^* just in case there is a word $w \in \Sigma^*$ such that $I = w\Sigma^*$, $I = \Sigma^* w$, or $I = \Sigma^* w \Sigma^*$ respectively. Note this is just like the definition of ideal in a monoid (Howie, 1995) except we are restricting the reference to a singleton set containing particular word w. Now let J_w denote the two-sided word ideal $\Sigma^* w \Sigma^*$. Then for $k \in \mathbb{N}$, the kth section of J_w is $\Sigma^k w \Sigma^* \subsetneq J_w$, which is denoted $J_{w,k}$.

Lemma 29. *For $i, j \in \mathbb{N}$ and where $|w| = l$,*

$$d(J_{w,i}, J_{w,j}) = 2\alpha^{-l}. \tag{10}$$

We can also compute the norm of $J_{w,i}$ when $|w| = l$:

$$\begin{aligned}
\|J_{w,i}\| &= \limsup_{k\to\infty} \frac{|\Sigma^i w \Sigma^{<k-i-l}|}{|\Sigma^{<k}|} \\
&= \limsup_{k\to\infty} \frac{\sum_{s=0}^{k-i-l-1} \alpha^{i+s}}{|\Sigma^{<k}|} \\
&= \limsup_{k\to\infty} \frac{\frac{\alpha^{k-l}-\alpha^i}{\alpha-1}}{\frac{\alpha^k-1}{\alpha-1}} \\
&= \limsup_{k\to\infty} \alpha^{-l} \left[\frac{\alpha^k - 1}{\alpha^k - 1} - \frac{\alpha^{i+l} - 1}{\alpha^k - 1}\right] \\
&= \alpha^{-l}. \qquad (11)
\end{aligned}$$

From Lemma 29 and the above calculation, taking $\mathsf{J}_{w,i} \stackrel{\text{def}}{=} \eta(J_{w,i})$, the sequence $\{\mathsf{J}_{w,i}\}_{i\in\mathbb{N}}$ is such that no subsequence can converge, yet every language in each point of the sequence has the same norm.

Lemma 30. *The Besicovitch quotient space $\mathcal{Q}_\zeta$ is not compact.*

Proof. It is sufficient to display an infinite sequence of languages belonging to distinct $\sim$ equivalence classes separated from each other by a distance greater than some fixed $\epsilon > 0$. Then the η-images of these languages will form an infinite sequence in $\mathcal{Q}_\zeta$ which has no convergent subsequence.

To this end, consider the language sequence $\mathbf{J}_a \stackrel{\text{def}}{=} \{J_{a,i}\}_{i\in\mathbb{N}}$ where $a \in \Sigma$. Two distinct terms $J_{a,i}$ and $J_{a,j}$ are at distance $2\alpha^{-1}$, from the previous lemma, so consider the sequence $\mathbf{L} = \{\mathsf{L}_i\}_{i\in\mathbb{N}}$, where $J_{a,i} \in \mathsf{L}_i$ for all $i \in \mathbb{N}$. By Corollary 25, there is no convergent subsequence of $\mathbf{L}$, since $\mathbf{d}(\mathsf{L}_i, \mathsf{L}_j) > \alpha^{-1}$ if $i \neq j$. □

Since sequential compactness is not defined in the pseudo-metric language space, we exhibit the following result to clear up any remaining doubts about compactness there.

Corollary 31. *The metric $\mathbf{d}$ is not complete.*

Proof. (Outline) It suffices to exhibit a sequence of points which are Cauchy convergent in $\mathcal{Q}_\zeta$, but which do not converge to any point in $\mathcal{Q}_\zeta$. We then produce a sequence, Cauchy in $\mathcal{Q}$, but containing the non-convergent sequence $\mathbf{L}$ from the proof of Lemma 30 as a subsequence.

□

Corollary 32. *A language space is not compact under the Besicovitch topology.*

Proof. Let $\mathcal{O}$ be an open cover of 2^{Σ^*} defined by

$$\mathcal{O} = \{\{M : d(L, M) < \alpha^{-1}\} : L \in 2^{\Sigma^*}\}.$$

We then have by Lemma 30 that any finite subset of $\mathcal{O}$ contains at most finitely many languages in $\mathbf{J}_a$. Therefore $\mathcal{O}$ has no finite subcover. □

While establishing noncompactness has been important, it will also be useful to establish a relation to a known compact space. This is the subject of the next subsection.

3.3 A second lifting of the quotient space

To obtain a compact space for exploring the most general features of the Besicovitch topology on language spaces, we define the language norm $\|\cdot\|_\zeta$ as a quotient map from $\mathcal{Q}_\zeta$ into $[0,1]$. This will result in a total of three spaces: the non-T_1 language space under the topology induced by Besicovitch distance, the quotient space topologized by the metric quotient topology, and a compact upper quotient space with a well-known topology. We proceed as with the definition of $\mathcal{Q}_\zeta$ by defining an equivalence relation, the equivalence classes, and the quotient map which takes points in $\mathcal{Q}_\zeta$ to their equivalence classes. We call the collection of equivalence classes the *upper Besicovitch quotient space,* denoted $\mathcal{N}_\zeta$. We ultimately show that the topological space $\mathcal{N}_\zeta$ under the quotient topology is homeomorphic to the unit interval.

Take points $\mathsf{L}, \mathsf{M} \in \mathcal{Q}_\zeta$. Let $\mathsf{L} \equiv_\zeta \mathsf{M}$ if $\|L\|_\zeta = \|M\|_\zeta$ for all $L \in \mathsf{L}$ and $M \in \mathsf{M}$; let $\langle \mathsf{L} \rangle_\zeta = \{\mathsf{M} \in \mathcal{Q}_\zeta : \mathsf{M} \equiv_\zeta \mathsf{L}\}$, and denote by $\mathcal{N}_\zeta$ the collection $\{\langle \mathsf{L} \rangle_\zeta : \mathsf{L} \in \mathcal{Q}_\zeta\}$, write elements of $\mathcal{N}_\zeta$ in calligraphy font $\mathcal{L}, \mathcal{M}, \mathcal{N}, \ldots$, and denote collections of such elements in corresponding bold letters $\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{N}}, \ldots$; let κ be the map from $\mathcal{Q}_\zeta$ to $\mathcal{N}_\zeta$ which takes L to its equivalence class $\langle \mathsf{L} \rangle_\zeta$. Finally, for $r \in [0,1]$, let $\mathbf{r}_\zeta$ denote $\{\mathsf{L} \in \mathcal{Q}_\zeta : \|L\|_\zeta = r \ \forall L \in \mathsf{L}\}$.

Remark 33. *It is obvious that $\equiv$ is an equivalence relation. Moreover, the quotient map κ is well-defined, by Corollary 26. Since $\mathbf{r}_\zeta = \langle \mathsf{M} \rangle_\zeta$ for each $\mathsf{M} \in \mathbf{r}_\zeta$, this implies by Remark 12 that $\mathbf{r}_\zeta = \mathcal{M}$ for precisely one $\mathcal{M} \in \mathcal{N}_\zeta$.*

We next equip the upper quotient space with a metric. Let the distance function $\rho : \mathcal{N}_\zeta \times \mathcal{N}_\zeta \to [0,1]$ be defined such that, if $\mathcal{L} = \mathbf{r}_\zeta$ and $\mathcal{M} = \mathbf{s}_\zeta$ for some $r, s \in [0,1]$, then $\rho(\mathcal{L}, \mathcal{M}) = |r - s|$ as a metric on $\mathcal{N}_\zeta$. The collection $\mathbf{U}$ of basis sets under the induced topology equals:

$$\{\{\boldsymbol{\mathcal{L}} \subset \mathcal{N}_\zeta : \mathbf{r}_\zeta \in \boldsymbol{\mathcal{L}} \text{ if } |r-s| < \epsilon > 0\} : s \in [0,1]\}. \tag{12}$$

Remark 34. *The set $\mathbf{U}$ is apparently equivalent to the subset topology on the unit interval. To wit, there is a homeomorphism between $\mathcal{N}_\zeta$ and $[0,1]$ if the function ρ induces the quotient topology on $\mathcal{N}_\zeta$.*

We continue to abuse the notation as was done with languages and the quotient space, and write $L \in \mathbf{r}_\zeta$ or equivalently $L \in \mathcal{L}$ to mean language L is found in points of the equivalence class $\mathbf{r}_\zeta$. We write $\mathbf{L} \subseteq \mathbf{r}_\zeta$ to mean that each language in the class $\mathbf{L}$ is in a point (not necessarily all in the same point) in the equivalence class $\mathbf{r}_\zeta$. We write $\mathbf{L} \subseteq \boldsymbol{\mathcal{L}}$ to mean that the image $\kappa(\eta(\mathbf{L}))$ is a subset of the collection of elements $\boldsymbol{\mathcal{L}} \subseteq \mathcal{N}_\zeta$. We will next show that, with exactly two exceptions, $\mathbf{r}_\zeta$ is always an uncountable subset of $\mathcal{Q}_\zeta$.

Lemma 35. *The $\equiv$ equivalence classes $\mathbf{0}_\zeta$ and $\mathbf{1}_\zeta$ are singletons in $\mathcal{Q}_\zeta$.*

Proof. The $\equiv$ class $\mathbf{0}_\zeta$ contains only the $\sim$ class $\eta(\emptyset)$, since $\|L\| = d(L, \emptyset)$. Thus, $L \in \mathbf{0}_\zeta$ implies $d(L, \emptyset) = 0$, which implies $L \sim \emptyset$.

On the other hand, suppose languages L and M and points L and M are such that $L \in \mathsf{L}$ and $M \in \mathsf{M}$ and $\mathsf{L}, \mathsf{M} \in \mathbf{1}_\zeta$. By Remark 12, $\|\neg L\| = \|\neg M\| = 0$, which we have just seen means $\neg L \sim \neg M$. But since $L \backslash M = \neg M \backslash \neg L$, it is true that $L \triangle M = \neg L \triangle \neg M$. Therefore $d(L, M) = \|L \triangle M\| = \|\neg L \triangle \neg M\| = d(\neg L, \neg M) = 0$. Hence, $L \sim M$. Since L, M were arbitrary, it follows that $\mathsf{L} = \mathsf{M}$ and that $\mathbf{1}_\zeta$ contains just a single point, viz. the equivalence class $\eta(\Sigma^*)$. □

Since $\mathbf{1}$ is a singleton, given a point L there is exactly one point in $\mathcal{Q}_\zeta$ at distance 1. If $\mathsf{L}, \mathsf{M} \in \mathcal{Q}_\zeta$ and $\mathbf{d}(\mathsf{L}, \mathsf{M}) = 1$, then points L, M will be called *antipodes,* which we denote as $\mathsf{L} = \neg\mathsf{M}$.

Lemma 36. *Every point $\mathsf{L} \in \mathcal{Q}_\zeta$ has a unique antipode in the Besicovitch quotient space.*

Proof. From Corollary 25 this is the same as claiming that, if two languages are at distance 1 from the same language $L \in \mathsf{L}$, then they are $\sim$-equivalent. But this is a consequence of the identity

$$(L \triangle M_1) \triangle (L \triangle M_2) = M_1 \triangle M_2. \tag{13}$$

We can show this because if $d(L, M_1) = 1$ and $d(L, M_2) = 1$, it follows that $L \triangle M_1$ and $L \triangle M_2$ are in $\mathbf{1}$ (from Def. 11), implying by Lemma 35 that $d(L \triangle M_1, L \triangle M_2) = 0$, requiring $M_1 \sim M_2$. □

Corollary 37. *For $L \in 2^{\Sigma^*}$, $\|\neg L\| = 1 - \|L\|$.*

In addition we note that $L \in \mathbf{0}$ iff $\neg L \in \mathbf{1}$, and also that $\langle \neg \mathsf{L} \rangle = \langle \mathsf{L} \rangle$ if and only if $\|L\| = \frac{1}{2}$ for any language $L \in \mathsf{L}$.

For each point $\mathsf{L} \in \mathfrak{Q}_\zeta$, the L-rotation of point $\mathsf{M} \in \mathfrak{Q}_\zeta$, denoted $\eta_{\mathsf{L}}(\mathsf{M})$, is defined as the point $\eta(L \triangle M)$ for some language $L \in \mathsf{L}$. The L-rotation of the Besicovitch quotient space, denoted $\mathfrak{Q}_\zeta^{\Sigma,\mathsf{L}}$, is then the collection $\{\eta_{\mathsf{L}}(\mathsf{M}) : \mathsf{M} \in \mathfrak{Q}_\zeta\}$. The L-rotation of the $\equiv$-equivalence class $\mathbf{r}$, denoted $\mathbf{r}_{\mathsf{L}}$, is defined as the set $\{M \in \mathfrak{Q}_\zeta : \mathbf{d}(\mathsf{M}, \mathsf{L}) = r\}$. The L-rotation of the upper Besicovitch quotient space, meaning the collection $\{\mathbf{r}_{\mathsf{L}} : r \in [0, 1]\}$, will be denoted $\mathcal{N}_{\zeta,\mathsf{L}}$.

Lemma 38. *$\mathfrak{Q}_\zeta^{\Sigma,\mathsf{L}}$ is equivalent as a set to $\mathfrak{Q}_\zeta$, and L-rotation is a bijection of the quotient space onto itself. Moreover, $\mathcal{N}_{\zeta,\mathsf{L}}$ is a bijection with $\mathcal{N}_\zeta$.*

There are uncountably many $\equiv$ equivalence classes, because the norm $\|\cdot\|$ is surjective onto the unit interval. In addition, we now proceed to show that no open set in $\mathfrak{Q}_\zeta$ is contained in a single $\equiv$ equivalence class. This is the essential condition for the proof that ρ is the quotient of $\mathbf{d}$. We begin with a straightforward proposition.

Proposition 39. *For $L \in 2^{\Sigma^*}$ where $\|L\| = r$ and $0 \leq s \leq r (\leq 1)$, there exists language $M \subset L$ such that $\|M\| = s$.*

Proof. For $s = 0$, let $M = \emptyset$, Q.E.D. For $s = r$, let $M = L$, Q.E.D. Now assume $s \in (0, r)$. Note that $s/r > 0$; form language sequence $\mathbf{L} = \{L \cap \Sigma^i\}_{i \in \mathbb{N}}$, and using this define the integer sequence $\{m_i\}_{i \in \mathbb{N}}$ such that

$$m_i = \lfloor (s/r) | (L \cap \Sigma^i) | \rfloor. \tag{14}$$

There exists a language sequence $\{M_i\}_{i \in \mathbb{N}}$ such that $M_i \subseteq L \cap \Sigma^i$ and $|M_i| = m_i$. Then we can calculate that $0 \leq (s/r)|(L \cap \Sigma^{<k})| = |(M \cap \Sigma^{<k})| < k$, so $\|M\| = (s/r)\|L\| = s$, and $M \subseteq L$; Q.E.D. □

Remark 40. *The above result can be reversed, in that if $0 \leq r \leq s \leq 1$, then for any language $L \in \mathbf{r}$ there exists language $M \supseteq L$ such that $M \in \mathbf{s}$. The target language is L in case $0 = r = s$, Σ^* in case $s = 1$, and in case $s \in (0, 1)$ may be constructed as in Proposition 39 by inverting the fractions in (14) et seq.*

Lemma 41. *No open set in $\mathfrak{Q}_\zeta$ is a subset of a $\equiv$ equivalence class.*

Proof. Since $\mathfrak{Q}_\zeta$ is perfect, this follows for the classes $\mathbf{0}_\zeta$ and $\mathbf{1}_\zeta$ directly from Lemmas 35 and 28. Otherwise, suppose $L \in 2^{\Sigma^*}$ and $\mathsf{L} \in \mathfrak{Q}_\zeta$ such that $L \in \mathsf{L} \in \mathbf{r}$. For any open set $\mathbf{L} \subset \mathfrak{Q}_\zeta$ containing L, there is a number $\epsilon' > 0$ such that $\mathbf{d}(\mathsf{L}, \mathsf{M}) < \epsilon'$ implies $\mathsf{M} \in \mathbf{L}$.

It is sufficient to exhibit a language $M \in \mathsf{M}$ such that $\|M\| \neq \|L\|$ and $d(L, M) < \epsilon'$. Let $\epsilon = \min\{r/2, \epsilon'/2\}$. Note that $\epsilon' > \epsilon > 0$. Our selection of ϵ guarantees the following: $0 < \epsilon < r \leq 1$, which implies that

$$0 < r - \epsilon < r \tag{15}$$

Then by Proposition 39 there is a language $M \subset L$ such that $\|M\| = r - \epsilon$. But since $r - \epsilon < r$, $\|M\| \neq \|L\|$. It also follows that $d(L, M) = \|L \triangle M\| = \|L \backslash M\|$. However, $\|L\| = \|M\| + \|L \backslash M\|$ from Corollary 14. Thus $d(L, M) = \|L\| - \|M\| = r - (r - \epsilon) = \epsilon$. □

Corollary 42. *If $\mathbf{L} \in \tau_\zeta$ is an open set in the Besicovitch topological space and language $L \in \mathbf{L}$, then there exists $\epsilon > 0$ such that for every real number*

$$r \in (\|L\| - \epsilon, \|L\| + \epsilon) \cap [0, 1]$$

there exists language $M \in \mathbf{L}$ such that $M \in \mathbf{r}$.

This corollary states that, under the Besicovitch topology, representatives of some continuous interval of norm values can be found in every open set in the language space. This means that, as was claimed in Remark 12(1), the language norm $\|\cdot\|_\zeta$ is a continuous map from 2^{Σ^*} onto $[0, 1]$.

Theorem 43. *The upper quotient space $\mathcal{N}_\zeta$ is homeomorphic to (and so essentially is) the unit interval $[0, 1]$.*

Ideals simplify exploration of the elements of the upper quotient space

Earlier we defined the (word) ideals of Σ^*. To elaborate on this, recall the earlier discussion of

r-simple languages (v. Def. 16), and consider the monoid ideals of Σ^*.

Lemma 44. *If real number $r \in [0,1]$, there exists a right ideal of Σ^* in $\mathbf{L}_r$.*

Proof. If $r = 1$ then $w = \lambda$ trivially satisfies the lemma. So we assume $r \in (0,1)$. Since by Def. 16 $0 \leq \check{r}_1 < \alpha$, there is a subset I_1 of Σ (actually, at least α subsets) such that $|I_1| = \check{r}_1$. Note from the definition of $\mathbf{L}_r$ that $\check{r}_k \leq r\alpha^k < \check{r}_k + 1$ for all $k \in \mathbb{N}$. Multiplying through by α gives the inequality

$$\check{r}_k\alpha \leq r\alpha^{k+1} < \check{r}_k\alpha + \alpha. \tag{16}$$

But for $k+1$ we have

$$\check{r}_{k+1} = \lfloor r\alpha^{k+1} \rfloor \leq r\alpha^{k+1} < \check{r}_{k+1} + 1. \tag{17}$$

Since all values are non-negative integers we can combine the preceding two equations to yield

$$\check{r}_k\alpha \leq \check{r}_{k+1} < \check{r}_k\alpha + \alpha. \tag{18}$$

It follows that $\check{r}_{k+1} = \check{r}_k\alpha + t_k$ for some $t_k \in \mathbb{N}$ such that $0 \leq t_k < \alpha$. Therefore for all $k \in \mathbb{N}$, $\check{r}_k\alpha \leq \alpha^{k+1} - \alpha$.

Thus there exists language $T_1 \subseteq \Sigma^2 \backslash I_1\Sigma$ such that $|T_1| = t_1$, so that $|I_1\Sigma \cup T_1| = \check{r}_2$. Set $I_2 = I_1\Sigma \cup T_1$. Continuing in this fashion, let T_k for each $k \in \mathbb{N}$ be a language such that $T_k \subseteq \Sigma^{k+1} \backslash I_k\Sigma$ and $|T_k| = t_k$. Finally, for $k \in \mathbb{N}$ define language $I \in 2^{\Sigma^*}$ such that $I \cap \Sigma^k = I_k$, which is to say let $I = \bigcup_{i\in\mathbb{N}} I_i$. Then by construction, $I \in \mathbf{L}_r$, and $w\Sigma^j \subseteq I$ for all $w \in I$ and every $j \in \mathbb{N}$. Thus $I\Sigma^* \subseteq I$. □

The preceding result provides further evidence that right ideals are ubiquitous in the Besicovitch topological space. We now develop our understanding of the ideals to comprehend the elements of the upper quotient space. We begin by extending the notion of "sections of a word ideal."

Definition 45. *An n-word ideal in the monoid Σ^* is a language J_F such that*

$$J_F = \Sigma^* w_1 \Sigma^* w_2 \ldots \Sigma^* w_n \Sigma^*$$

for some finite language $F = \{w_1, w_2, \ldots, w_n\}$ over Σ^*. Then $f_F = \sum_{i=1}^{n} |w_i|$ is the length of F. If $\mathbf{v} = (v_1, \ldots, v_n)$ is a vector over $\mathbb{N}^{1\times n}$, then the $\mathbf{v}$-section of J_F is denoted $J_{F,\mathbf{v}}$ and is the right ideal defined as:

$$J_{F,\mathbf{v}} = \Sigma^{v_1} w_1 \Sigma^{v_2} w_2 \Sigma^{v_3} \cdots \Sigma^{v_n} w_n \Sigma^*.$$

Lemma 46. *For every vector $\mathbf{v}$ over $\mathbb{N}^{1\times n}$ and every language F such that $|F| = n$, $\|J_{F,\mathbf{v}}\| = \alpha^{-f_F}$.*

Proof. Let $v_1 + v_2 + \ldots + v_n = S$. Then when $k \geq f_F + S$, $|J_{F,\mathbf{v}} \cap \Sigma^k| = \alpha^{k-f_F}$. Therefore,

$$\begin{aligned}
&\lim_{k\to\infty} \frac{\sum_{i=0}^{k-1} \alpha^{i-f_F} - |J_{F,\mathbf{v}} \cap \Sigma^{<k}|}{\sum_{i=0}^{k-1} \alpha^i} \\
&= \lim_{k\to\infty} \frac{\sum_{i=0}^{f_F+S-1} \alpha^{i-f_F}}{\sum_{i=0}^{k-1} \alpha^i} \\
&= \lim_{k\to\infty} \frac{\frac{\alpha^S-1}{\alpha-1}}{\frac{\alpha^k-1}{\alpha-1}} = 0,
\end{aligned}$$

which implies that $\|J_{F,\mathbf{v}}\| = \alpha^{-f_F}$. □

In addition to the above result, it is possible to extend the proofs of Lemmas 44 and 30 to the n-word ideals by induction. Taken together, these results tell us that points in the upper quotient space contain languages that "closely resemble" unions of sections of ideals of Σ^*, in the following sense: cardinality of sections of these languages (as word length increases) must approximate the cardinality of the unions of (sections of) ideals.

We conclude this section by showing that all $\equiv$ classes except $\mathbf{0}_\zeta$ and $\mathbf{1}_\zeta$ are uncountable.

Lemma 47. *For any real number $r \in (0,1)$, the element $\mathbf{r} \in \mathcal{N}_\zeta$ is uncountable.*

Proof. From Lemma 44, there is an r-simple language L. In fact, there exist at least two r-simple languages, since for each $r \in (0,1)$,

$$\begin{aligned}
0 \leq |L \cap \Sigma^k| &< \left[r + \frac{1-r}{2} \right] \alpha^k \\
&= \left(\frac{r+1}{2} \right) \alpha^k < r\alpha^k.
\end{aligned}$$

This means that for $k \in \mathbb{N}$ there exists a subset of $\Sigma^k \backslash L = \neg(L \cap \Sigma^k)$ consisting of the lesser of either $\lfloor r\alpha^k \rfloor$ or $\left\lfloor \left(\frac{1-r}{2}\right) \alpha^k \right\rfloor$ words, and there exists a subset of $L \cap \Sigma^k$ consisting of the same number

of words. This means there exists an r-simple language at distance $s = \min\{2r, 1-r\}$ from L. We now construct this language in the following way: let $t_k = \min\left\{\lfloor r\alpha^k \rfloor, \left\lfloor \left(\frac{1-r}{2}\right)\alpha^k \right\rfloor\right\}$; let T_k be a language such that $|T_k| = t_k$ and $T_k \subseteq \neg(L \cap \Sigma^k)$, which is possible since $|\neg(L \cap \Sigma^k)| \geq 2t_k$; and let $F_k \subseteq L$ be such that $|F_k| = t_k$, which is possible since $t_k \leq |L \cap \Sigma^k|$. Let $T = \bigcup_{i\in\mathbb{N}} T_i$, $F = \bigcup_{i\in\mathbb{N}} F_i$, and let $N = L \backslash F$. Then language $L' \stackrel{\text{def}}{=} N \cup T$ is the language formed by exchanging t_k words in L for t_k words in $\neg L$. Thus the number of words in $L \triangle^k L'$ is $2t_k = s\alpha^k$ for each $k \in \mathbb{N}$. Hence, $d(L, L') = s$ and, since L and L' contain the same number of words of each length, they have the same norm. Since L is r-simple, so is L'.

For all $t \in \mathbb{R}$ such that $0 \leq t \leq s$, since $s \leq r$ there exists language $F' \subseteq F \subseteq L$ such that $\|F'\| = t/2$, and there exists language $T' \subseteq T = L' \cap \neg L$ such that $\|T'\| = t/2$ (by Proposition 39). Then it can be shown that if language $L_t \stackrel{\text{def}}{=} (L \backslash F') \cup T'$ is such that

$$L_t = (L \backslash F) \cup (F \backslash F') \cup T' = N \cup (F \backslash F') \cup T',$$

so $L_t \triangle L' = (T \backslash T') \cup (F \backslash F')$. Thus $d(L_t, L') = s - t$. □

3.4 The Chomsky hierarchy

In this final section we show a few results which relate our Besicovitch topologies to the classical language classes.

The finite and locally testable languages are not dense

A major inadequacy of the Cantor topology was the density of the finite languages. By contrast, these are confined to a single $\sim$-equivalence class in the Besicovitch topology.

Lemma 48. *The finite languages are all in* $\mathbf{0}_\zeta$.

Proof. If language L is finite, there exists $N \in \mathbb{N}$ such that $n > N$ implies $L \cap \Sigma^n = \emptyset$, and hence also that $|L \triangle^n \emptyset| = 0$. □

This naturally leads to the question, addressed presently, what happens if the description of an infinite language is entirely finitary?

We first remind the reader that a language L is *locally testable* just in case there is a fixed integer k (called a window length) and a proper subset $F \subsetneq \Sigma^k$ such that, if every factor of word w of length k is in F then $w \in L$. The important thing about the locally testable family is that the membership question "Is $w \in L$?" is decidable by inspecting subsequent k-length factors of w. We next define a larger class of "generally testable" languages with the property that every locally testable language is a subset of some generally testable language.

Definition 49. *A language L is* generally testable *if there exists a window length $n \in \mathbb{N}$ and a set of permitted factors $S \subseteq \Sigma^n$, where $L = S^*\Sigma^{<n}$.*

From this definition we see that word $w \in L$ if and only if $w \in \Sigma^{<n}$ or w can be written $u_1u_2\cdots u_tv$, where $u_i \in S$ for all $i \in \mathbb{N}_t$ and $v \in \Sigma^{<n}$. It is interesting that the size of a generally testable language is not really limited, but yet we have the following result.

Lemma 50. *Every generally testable language in 2^{Σ^*} is in $\mathbf{0}_\zeta$ with the exception of Σ^*, which is in $\mathbf{1}_\zeta$.*

Proof. (Outline) Let the permitted factors of a word in L be $S \subseteq \Sigma^n$. If $|S| = s$, suppose $s = \alpha^n$. But then $S = \Sigma^n$, $L = \Sigma^*$, and therefore $L \in \mathbf{1}_\zeta$.

On the other hand, if $s < \alpha^n$, and word $w \in L$, there exist unique non-negative integers q and r, such that $|w| = nq + r$ and $0 \leq r < n$, and words $u_1, u_2, \ldots, u_q$ in S, and word v in Σ^r such that $w = u_1u_2\ldots u_qv$We deduce that $|L \cap \Sigma^{|w|}| = s^q\alpha^r$.

We can therefore easily see that the proportion of the number of words $L^{<i}$ to those in $\Sigma^{<k}$ is maximized at word lengths where $q = n-1$, i.e., where $i = nk + n - 1$. We conclude the following:

$$\|L\| \leq \limsup_{k\to\infty} \frac{\sum i = 0^k s^i|(|\Sigma^{<n}}{|\Sigma^{<kn+n}|}. \qquad (19)$$

By our assumption, $s \leq \alpha^n - 1$. Straightforward calculation shows the right side of the above equation tends to zero, because it is bound above by

$$\limsup_{k\to\infty} \frac{1}{2}\frac{(\alpha^n - 1)^k}{(\alpha^n)^k}.$$

Thus, $L \in \mathbf{0}_\zeta$.

□

Corollary 51. *Every locally testable language belongs to* $\mathbf{0}_\zeta$.

Proof. Suppose L is a locally testable language over Σ with window length n and permitted factors $S \subsetneq \Sigma^n$. Consider the generally testable language L' with the same window length and the same permitted factors as locally testable language L. Then $L \subseteq L'$ and, by the properties of a language norm, $\|L\| \leq \|L'\|$; meanwhile $\|L'\| = 0$ from the preceding lemma. □

Regular languages are dense in the upper quotient space

We have now seen that all finite and locally testable languages belong to $\mathbf{0}_\zeta$. On the other hand:

Lemma 52. *Regular languages are dense in the upper quotient space* $\mathcal{N}_\zeta$.

Proof. Let $r \in [0, 1]$. The claim is that for all $\epsilon > 0$ there exists a regular language L such that $|\|L\| - r| < \epsilon$. If $\epsilon \geq \min\{r, 1 - r\}$, either $\emptyset$ or Σ^* satisfies the claim, Q.E.D. So we assume that $\epsilon < \min\{r, 1 - r\}$. Then $r < r + \epsilon < 1$. Let integers n and q be such that $r < q\alpha^{-n} \leq r + \epsilon \leq (q + 1)\alpha^{-n}$, and $0 < q < \alpha^n$. From this we have

$$0 < q\alpha^{-n} - r < \epsilon. \tag{20}$$

Let language $S_\epsilon \subseteq \Sigma^n$ have cardinality q. Consider the right ideal $S_\epsilon\Sigma^*$, which is a disjoint union of the q right word ideals $w\Sigma^*$ with $w \in S_\epsilon$. Note that each of these is a 1-word ideal section $J_{F,\mathbf{v}}$, where $F = \{w\}$ for $w \in S_\epsilon$ and $\mathbf{v} = (0)$. Therefore by Lemmas 13 and 46,

$$\begin{aligned} \|S_\epsilon\Sigma^*\| &= \sum_{w \in S_\epsilon} \|w\Sigma^*\| \\ &= q\alpha^{-n} \end{aligned}$$

From (20) this means that $|\|S_\epsilon\Sigma^*\| - r| < \epsilon$ as required. Finally, by the Myhill-Nerode Theorem (Nerode, 1958) $S_\epsilon\Sigma^*$ is a regular language, since all but finitely many words in $S_\epsilon\Sigma^*$ can be followed by Σ^*. □

This means that the linear, context-free, context-sensitive, and recursively enumerable languages are all dense in the upper quotient space. We still do not know where all these families lie in the lower Besicovitch topological spaces, but we conjecture that the regular languages are indeed also dense in the Besicovitch topology $(2^{\Sigma^*}, \tau_\zeta)$.

Non-r.e. languages are dense in both quotient spaces

We can show fairly simply that the non-recursively enumerable languages are ubiquitous in the Besicovitch topological spaces. Because d_ζ is a strict pseudo-metric, the $\sim$ equivalence classes are uncountable. We present the following without their (uncomplicated) proofs due to space limitations.

Lemma 53. *The single element of the class* $\mathbf{0}_\zeta$ *is uncountable in* 2^{Σ^*} *and contains a non-r.e. language.*

Corollary 54. *Every* $\sim$ *equivalence class contains a non-r.e. language.*

4 Conclusion

We have attempted to improve upon previous definitions of *distance* between languages in a language space. After considering previous work by Vianu (1977) which defined a language distance using the density of their symmetric set difference, we progressed to a new adaptation of a pseudometric inspired by Besicovitch (1932). In a language space, the Besicovitch pseudometric was developed which is essentially the upper density of the set-difference between languages. By lifting to the quotient space $\mathcal{Q}_\zeta$ using Besicovitch equivalence, a natural metric topology was developed and shown to be perfect but not compact. Another step of lifting brought us a compact "upper" quotient space $\mathcal{N}_\zeta$ homeomorphic to the unit interval. The ideals of this upper space were studied, also invoking the notion of *word ideal* defined herein. In the last section it was shown that neither the finite nor locally testable languages are dense in $\mathcal{N}_\zeta$. Finally, the regular languages were shown to be dense in $\mathcal{N}_\zeta$, and the non-r.e. languages were shown to be dense in both $\mathcal{Q}_\zeta$ and $\mathcal{N}_\zeta$.

References

J. Berstel. 1973. Sur la densité asymptotique de langages formels. In *International Colloquium on Automata, Languages and Programming (ICALP, 1972)*, pages 345–358. North-Holland.

A. S. Besicovitch. 1932. *Almost Periodic Functions*. The University Press.

V. G. Bodnarchŭk. 1965. The metrical space of events, part I. *Kibernetika*, 1(1):24–27.

C. S. Calude, H. Jürgensen, and L. Staiger. 2009. Topology on words. *Theoretical Computer Science*, 410:2323–2335.

G. Cattaneo, E. Formenti, L. Margara, and J. Mazoyer. 1997. A shift-invariant metric on $s^{\mathbb{Z}}$ inducing a non-trivial topology. In I. Privara and P. Rusika, editors, *Mathematical Foundations of Computer Science 1997*, volume 1295 of *LNCS*, pages 179–188. Springer-Verlag.

A. Dincă. 1976. The metric properties on the semigroups and the languages. In A. Mazurkiewicz, editor, *Mathematical Foundations of Computer Science 1976*, volume 45 of *LNCS*, pages 260–264. Springer-Verlag.

M. Gehrke. 2009. Stone duality and the recognisable languages over an algebra. In A. Kurz, M. Lenisa, and A. Tarlecki, editors, *Algebra and coalgebra in computer science*, volume 5728 of *LNAI*, pages 236–250. Springer.

D. Genova and N. Jonoska. 2006. Topological properties of forbidding-enforcing systems. *Journal of Automata, Languages and Combinatorics*, 11(4):375–398.

J. M. Howie. 1995. *Fundamentals of Semigroup Theory*. Oxford University Press.

J. Kozik. 2005. Conditional densities of regular languages. *Electronic Notes in Theoretical Computer Science*, 14.

J. Kozik. 2006. *Decidability of relative density in Chomsky hierarchy of languages*. Ph.D. thesis, Jagiellonian University, Cracow, Poland.

S. Marcus. 1966. *Introduction mathématique à la linguistique structurale*. Dunod.

S. Marcus. 1967. *Algebraic Linguistics; Analytical Models*. Academic Press.

E. Nelson. 1980. Categorical and topological aspects of formal languages. *Mathematical Systems Theory*, 13:255–273.

A. Nerode. 1958. Linear automaton transformations. *Proceedings of the American Mathematical Society*, 9(4):541–544.

A. Salomaa and M. Soittola. 1978. *Automata-theoretic aspects of formal power series*. Springer, Berlin.

V. Vianu. 1977. The Bodnarchŭk metric space of languages and the topology of the learning space. In J. Gruska, editor, *Mathematical Foundations of Computer Science 1977*, volume 53 of *LNCS*, pages 537–542. Springer-Verlag.

H. Walter. 1975. Topologies on formal languages. *Mathematical Systems Theory*, 9:142–158.

S. Yu. 1997. Regular languages. In G. Rozenberg and A. Salomaa, editors, *Handbook of formal languages*, volume 1, pages 41–110. Springer, Berlin.