Beyond n-grams

(1) Strategy thus far:
- Try to infer boundaries by looking extremely locally
- For each pair \(x, y\), ask “How surprising is it to see a \(y\) immediately after an \(x\)?”
  - If not very surprising, assume they are part of a cohesive unit
  - If sufficiently surprising, assume there is a boundary between \(x\) and \(y\)
- Effectiveness of this very simple strategy: not spectacular
  - Might provide a good start; useful as part of a bootstrapping strategy
  - Better for some tasks (and some languages?) than for others

(2) Where from here: explore some other strategies for finding word- and morpheme-like chunks

Successor counts: Harris (1955)

(3) The problem with n-grams:
- In general in English, the sequence [dv] isn't all that common; the transitional probability from \(d\) to \(v\) is fairly low. Thus, a boundary is fairly likely: [d#v]
- In the sequence […ardv…], however, the [v] is not at all surprising! In fact, it's virtually the only thing you would expect to hear after an [ard]
- Upshot: we could probably do better if we pay attention to longer strings of material, of varying length depending on what the material is.

(4) Leading idea: at each spot in the string, calculate how many different things could possibly come next—e.g., for the sentence *Aardvarks are cool* [ardvarksarkul]
- After [ar]: many different sentences begin with [ar], so there are many possible continuations.
  - k: *Arkansas is muggy in the summer.*
  - b: *Arborvitaes grows well in this climate.*
  - j: *Are you going?*
  - d: *Are they here yet?*
  - æ: *Are animals smarter than people?*
    - etc… (all English phonemes except [Z] and [N])
- After [ard]: just a few possibilities
  - a: *Ardent fans will be disappointed.*
  - r: *Ardor suits him.*
  - v: *Aardvarks subsist on insects and grubs.*
- Claim: it's plausible to at least consider the possibility of a boundary after [ar], but once you've heard [ard], you know the next few phonemes will have to be part of the same word.

(5) Some terminology:
- Prefix: the initial portion of a string, up to some point \(n\)
- Successor: the segment immediately following the prefix

| a r d v a r k s a r k u l | 1 2 3 4 5 6 7 8 9 10 11 12 13 |
(6) Successor count: for a prefix of length $n$, how many possible successors are there in the language?

- Intuition: spots with high successor counts are likely to be boundaries. (Why?)

(7) An more detailed example of how we might calculate and use successor counts (Harris 1955)

He's clever

<table>
<thead>
<tr>
<th>h</th>
<th>i</th>
<th>y</th>
<th>z</th>
<th>k</th>
<th>l</th>
<th>v</th>
<th>c</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

(8) Starting from the left: $n = 1$, Prefix = [h]

Possible successors to an initial [h] in English: 9 total[

- [i] His ship's in.  
  - hizSipsin
- [e] Hell, what's the use?  
  - helhw@tsD@yus
- [æ] Had to, sorry.  
  - hadtuwsari
- [a] Hot coffee.  
  - hatkafi
- [o] Her timing's off.  
  - h@rtaymiNzaf
- [ɔ] Hunting is a dumb thing to do.  
  - hɔntiNzad@mTiNtuduw
- [u] Hope for the best.  
  - howpfowrD@best
- [ʌ] Hook and ladder company.  
  - huk@nd1ad@rk@mp@ni
- [ɪ] Who is it?  
  - huwizit
- [ɔ] Humans act like simians.  
  - hyuwm@nz@klayksimi@nz

(9) Moving to the right: $n = 2$, Prefix = [hi]

Possible continuations to initial [hi]: 14 total

- [p] Hip high in water.  
  - hiphayinwat@r
- [t] Hit it back.  
  - hititbæk
- [k] Hickory nuts are still available.  
  - hik@rin@tsarstil@veyl@b@l
- [d] Hidden meanings were discovered.  
  - hid@mmiyNiNzw@rdisk@v@rd
- [g] Higginbotham is his name.  
  - higinbat@mizhizneym
- [ɔ] Hither, thither and yon.  
  - hiD@rTiD@rændyan
- [s] History will repeat itself.  
  - hist@riywilripiytitsel
- [ʃ] Hitch a ride with John.  
  - hiC@raydwiTJan
- [z] His work here is done.  
  - hizw@rkiyrizd@n
- [l] Hill-climbing is not my favorite activity.  
  - hillklaymiNiznatmayfeyvritaktivitiy
- [m] Himalayan goats are stubborn.  
  - him@ley@ngowtsarst@b@rn
- [n] Hindrances were everywhere.  
  - hindr@ns@sw@revriyhweyr
- [h] He has a coat on.  
  - hihæz@kowtan
- [y] He's a good friend.  
  - hiyz@gudfrend

(10) Moving to the right: $n = 3$, Prefix = [hiy]

Possible continuations to initial [hiy]: 29 total

---

1The phonemic transcriptions that we’ll use here are a bit weird, particularly for the vowels. I’m using the same system that Harris (1955) used, so that the numbers come out roughly equivalent to the ones in his paper.
He played volleyball.  

Heat is needed.  

He cooked dinner.  

He bought a car.  

Heed my words.  

He goes there often.  

He followed me home.  

Heave ho!  

Heathens were numerous.  

He scribbled furiously.  

He shot himself in the foot.  

He's a good friend.  

He chopped broccoli.  

He thought better of it.  

Heard a bird.  

He went along with it.  

He yapped all night.  

He isn't alone.  

He estimated it.  

He asked me about it.  

He offered me a smoke.  

He allowed it.  

He opened the door.  

He u???

Continued over the whole utterance:

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Successor count</th>
<th>Possible successors</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>9</td>
<td>w,y,i,e,æ,a,æ,a,o,u</td>
</tr>
<tr>
<td>hi</td>
<td>14</td>
<td>p,t,k,d,g,D,s,z,C,l,m,n,h,y</td>
</tr>
<tr>
<td>hiy</td>
<td>29</td>
<td>p,t,k,b,d,g,f,T,v,D,s,z,S,C,J,l,r,m,n,h,w,y,i,e,æ,a,æ,a,o,u</td>
</tr>
<tr>
<td>hiyz</td>
<td>29</td>
<td>p,t,k,b,d,g,f,T,v,D,s,S,C,J,l,r,m,n,h,w,y,i,e,æ,a,æ,a,o,u</td>
</tr>
<tr>
<td>hiyzk</td>
<td>11</td>
<td>l,r,w,y,i,e,æ,a,æ,a,o,u</td>
</tr>
<tr>
<td>hiyzkl</td>
<td>7</td>
<td>i,e,æ,a,æ,a,o,u,a</td>
</tr>
<tr>
<td>hiyzkle</td>
<td>1</td>
<td>v</td>
</tr>
<tr>
<td>hiyzklev</td>
<td>1</td>
<td>@</td>
</tr>
<tr>
<td>hiyzkle@</td>
<td>1</td>
<td>r</td>
</tr>
<tr>
<td>hiyzklev@r</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References