Using Electronic Dictionaries

Michael Cysouw
Philipps Universität Marburg

Experiences

- Digitising dictionaries: problems and solutions
- Using digital dictionaries for language comparison

Live or Legacy?

- **“Live” conversion**
  Integration of a dictionary-in-the-making into a comparative infrastructure
  (Problem: workflow compatibility)
- **“Legacy” conversion**
  Converting any fixed resource into a digitally accessible form
  (Problem: copyright)

Resources

- **“String” resources**
  each entry is a single string of characters with visual markup
  (“word file”, print dictionary)

- **“Tagged” resources**
  each entry consist of a set of tagged strings
  (“database”)

Conversion

- **Interpretation of string resources**
  (“word file → database”)
  complete interpretation of internal structure is highly laborious
- **Interpretation of tagged resources**
  (“connecting databases”)
  different traditions and different languages are full of idiosyncrasies that are difficult (impossible?) to match

mítyane ó áábímyeíhi. Tengo mucho temor por la enfermedad que viene.

abíhábi *onom.* 1. expresa que se prenden llamas de fuego. 2. expresa el estado de tener pintas redondas en la superficie.

aábo *abs.* insulto. ‖ acción de...

[aabo] *vt.* 1. poner trampa. *Áánu aabó ípakyééju.* El pone trampa en su represa (quebrada cerrada para que los peces no puedan pasar). 2. (fig.) insultar, ultrajar. *Tábyeebe oke aabó tátyájkiívá újtsiñe.* Mi sobrino me insultó diciéndome que mis piernas son muy delgadas.

áábojcátsi *abs.* insultos. *¿A úhdityúha tsáma teene áábojcátsi?* ¿Tú eres el que provocas los insultos? ‖ acción de...

[áábójcatsi] *vrec.* insultarse el uno al otro. *¿Íveekí ámuha máábócatsíhijcyá? ¡Íміáámére bo meíjcyaj!* ¿Por qué se insultan? ¡Vivan en armonía!

aabópi *abs.* estado de...

[aabópi] *ve.* ser insultante. *Tsaapi táñahbémudítyú aabópí.* Uno de mis hermanos es insultante.

[ábópí(h)] *adj.* insultante. *Tsaapi táñahbémudítyú ávyeta ábópí.* Uno de mis hermanos es muy insultante.

aabúcu *abs.* aguante, tolerancia, resistencia. ‖ acción de...

[aabúcu] *vt.* aguantar, soportar, tolerar, resistir. *Ííju aabúcú mítyane pádúúcuú.* El caballo aguanta mucho peso.

[aabúcu] *ve.* ser tolerante, ser resistente.

[ábúcú(h)] *adj.* tolerante, resistente. *Éje, eene tsíímene ábúcú tsivá ene piichúcoba.* Mira, ese niño resistente trae esa tremenda carga.

aabyúcu, aábyu *abs.* desenterramiento. ‖ acción de...

[aabyúcu, aabyu] *vt.* sacar, desenterrar algo. *Éíjyúu llihíyó aabyúcú ímyeemého.* Hace poco mi papá desenterró su masa de pijuayo (que había guardado).

ábyucúúve *abs.* efecto de...

[ábyúcuuve] *vi.* ser sacado lo que estaba metido en una cosa.

aca *part.* expresa duda. *¿Aca ure ú méénune?* ¿Lo has hecho solo?

aaca *conj.adv.* se refiere a una acción anterior. *Núhbadi tsá mítyane u íjcyáityuró; aaca tsá u chéméityuróne.* Si no hubieras estado mucho en el sol no te hubieras enfermado.

acádsi *onom.* expresa la acción de dejar de hacer algo. *¡Iijyévéné 'acádsi' u méénúcuhíjcyáné wáábyau u éjécunúne!* ¡No sueltes la soga a cada rato! *Ávyeta 'acádsi' néétune muha méwákímyeí.* Estamos trabajando de corrido sin tener tiempo para otra cosa.

acádsíh-acádsi *onom.* expresa que algo se suelta o se afloja poco a poco.

acádsihnécu *adv.* soltando instantáneamente. *Ávyeta aadi áákityé íañújú acádsihnécu.* Aquél se cayó y soltó instantáneamente su escopeta.

ácádsíjcaáyo, ácadsíjco *abs.* acción de...

[ácádsíjcaayo, ácádsijco] *vt.* 1. soltar, libertar, librar. 2. soltar, dejar caer. *Ú ácádsíjcaayó díwaajácuhéámí baávu.* Tú has dejado caer el libro al suelo.

ácadsííve, áhcadsíba *abs.* soltura; liber-

mítyane ó áábímyeíhi. Tengo mucho temor por la enfermedad que viene.

abíhábi *onom.* 1. expresa que se prenden llamas de fuego. 2. expresa el estado de tener pintas redondas en la superficie.

aábo *abs.* insulto. ‖ acción de...

[aabo] *vt.* 1. poner trampa. *Áánu aabó ípakyééju.* El pone trampa en su represa (quebrada cerrada para que los peces no puedan pasar). 2. (fig.) insultar, ultrajar. *Tábyeebe oke aabó tátyájkiívá újtsiñe.* Mi sobrino me insultó diciéndome que mis piernas son muy delgadas.

áábojcátsi *abs.* insultos. *¿A úhdityúha tsáma teene áábojcátsi?* ¿Tú eres el que provocas los insultos? ‖ acción de...

[áábójcatsi] *vrec.* insultarse el uno al otro. *¿Íveekí ámuha máábócatsíhijcyá? ¡Íміáámére bo meíjcyaj!* ¿Por qué se insultan? ¡Vivan en armonía!

aabópi *abs.* estado de...

[aabópi] *ve.* ser insultante. *Tsaapi táñahbémudítyú aabópí.* Uno de mis hermanos es insultante.

[ábópí(h)] *adj.* insultante. *Tsaapi táñahbémudítyú ávyeta ábópí.* Uno de mis hermanos es muy insultante.

aabúcu *abs.* aguante, tolerancia, resistencia. ‖ acción de...

[aabúcu] *vt.* aguantar, soportar, tolerar, resistir. *Ííju aabúcú mítyane pádúúcuú.* El caballo aguanta mucho peso.

[aabúcu] *ve.* ser tolerante, ser resistente.

[ábúcú(h)] *adj.* tolerante, resistente. *Éje, eene tsíímene ábúcú tsivá ene piichúcoba.* Mira, ese niño resistente trae esa tremenda carga.

aabyúcu, aábyu *abs.* desenterramiento. ‖ acción de...

[aabyúcu, aabyu] *vt.* sacar, desenterrar algo. *Éíjyúu llihíyó aabyúcú ímyeemého.* Hace poco mi papá desenterró su masa de pijuayo (que había guardado).

ábyucúúve *abs.* efecto de...

[ábyúcuuve] *vi.* ser sacado lo que estaba metido en una cosa.

aca *part.* expresa duda. *¿Aca ure ú méénune?* ¿Lo has hecho solo?

aaca *conj.adv.* se refiere a una acción anterior. *Núhbadi tsá mítyane u íjcyáityuró; aaca tsá u chéméityuróne.* Si no hubieras estado mucho en el sol no te hubieras enfermado.

acádsi *onom.* expresa la acción de dejar de hacer algo. *¡Iijyévéné 'acádsif u méénúcuhíjcyáné wáábyau u éjécunúne!* ¡No sueltes la soga a cada rato! *Ávyeta 'acádsif néétune muha méwákímyeí.* Estamos trabajando de corrido sin tener tiempo para otra cosa.

acádsíh-acádsi *onom.* expresa que algo se suelta o se afloja poco a poco.

acádsihnécu *adv.* soltando instantáneamente. *Ávyeta aadi áákityé íañújú acádsihnécu.* Aquél se cayó y soltó instantáneamente su escopeta.

ácádsíjcaáyo, ácadsíjco *abs.* acción de...

[ácádsíjcaayo, ácádsijco] *vt.* 1. soltar, libertar, librar. 2. soltar, dejar caer. *Ú ácádsíjcaayó díwaajácuhéámí baávu.* Tú has dejado caer el libro al suelo.

ácadsííve, áhcadsíba *abs.* soltura; liber-

mítyane ó áábímyeíhi. Tengo mucho temor por la enfermedad que viene.

abíhábi *onom.* 1. expresa que se prenden llamas de fuego. 2. expresa el estado de tener pintas redondas en la superficie.

aábo *abs.* insulto. || acción de...

[aabo] *vt.* 1. poner trampa. *Áánu aabó ípakyééju.* El pone trampa en su represa (quebrada cerrada para que los peces no puedan pasar). 2. (fig.) insultar, ultrajar. *Tábyeebe oke aabó tátyájkiívá újtsiñe.* Mi sobrino me insultó diciéndome que mis piernas son muy delgadas.

áábojcátsi *abs.* insultos. *¿A úhdityúha tsáma teene áábojcátsi?* ¿Tú eres el que provocas los insultos? || acción de...

[áábójcatsi] *vrec.* insultarse el uno al otro. *¿Íveekí ámuha máábócatsíhijcyá? ¡Ímiáámére bo meíjcyaj!* ¿Por qué se insultan? ¡Vivan en armonía!

aabópi *abs.* estado de...

[aabópi] *ve.* ser insultante. *Tsaapi táñahbémudítyú aabópí.* Uno de mis hermanos es insultante.

[ábópí(h)] *adj.* insultante. *Tsaapi táñahbémudítyú ávyeta ábópí.* Uno de mis hermanos es muy insultante.

aabúcu *abs.* aguante, tolerancia, resistencia. || acción de...

[aabúcu] *vt.* aguantar, soportar, tolerar, resistir. *Ííju aabúcú mítyane pádúúcuú.* El caballo aguanta mucho peso.

[aabúcu] *ve.* ser tolerante, ser resistente.

[ábúcú(h)] *adj.* tolerante, resistente. *Éje, eene tsíímene ábúcú tsivá eene piichúcoba.* Mira, ese niño resistente trae esa tremenda carga.

aabyúcu, aábyu *abs.* desenterramiento. || acción de...

[aabyúcu, aabyu] *vt.* sacar, desenterrar algo. *Éíjyúu llihíyó aabyúcú ímyeemého.* Hace poco mi papá desenterró su masa de pijuayo (que había guardado).

ábyucúúve *abs.* efecto de...

[ábyúcuuve] *vi.* ser sacado lo que estaba metido en una cosa.

aca *part.* expresa duda. *¿Aca ure ú méénune?* ¿Lo has hecho solo?

aaca *conj.adv.* se refiere a una acción anterior. *Núhbadi tsá mítyane u íjcyáítyuró; aaca tsá u chéméítyuróne.* Si no hubieras estado mucho en el sol no te hubieras enfermado.

acádsi *onom.* expresa la acción de dejar de hacer algo. *¡Iijyévéné 'acádsi' u méénúcuhíjcyáné wáábyau u éjécunúne!* ¡No sueltes la soga a cada rato! *Ávyeta 'acádsi' néétune muha méwákímyeí.* Estamos trabajando de corrido sin tener tiempo para otra cosa.

acádsíh-acádsi *onom.* expresa que algo se suelta o se afloja poco a poco.

acádsihnécu *adv.* soltando instantáneamente. *Ávyeta aadi áákityé íañújú acádsihnécu.* Aquél se cayó y soltó instantáneamente su escopeta.

ácádsíjcaáyo, ácadsíjco *abs.* acción de...

[ácádsíjcaayo, ácádsijco] *vt.* 1. soltar, libertar, librar. 2. soltar, dejar caer. *Ú ácádsíjcaayó díwaajácuhéámí baávu.* Tú has dejado caer el libro al suelo.

ácadsííve, áhcadsíba *abs.* soltura; liber-

ID	Entry	Page
Page 27, ID 5 Link by ID	**áábojcátsi** *abs.* insultos. *¿A úhdityúha tsáma teene áábojcátsi?* ¿Tú eres el que provocas los insultos? ‖ acción de... **[áábójca̲tsi]** *vrec.* insultarse el uno al otro. *¿Íveeki ámuha máábócatsihijcyá? ¡Ímiáámére bo meijcyaj!* ¿Por qué se insultan? ¡Vivan en armonía!	Page: 27, Column 1 View page

Annotations for main entry

Base text

Entry text:

```
áábojcátsi abs. insultos. ¿A úhdityúha tsáma teene áábojcátsi? ¿Tú eres el que provocas los insul
```

Python code

Type	Value	Start	End	Substring
dictinterpretation	head	0	13	áábojcátsi
dictinterpretation	pos	14	18	abs.
dictinterpretation	translation	18	27	insultos
dictinterpretation	example-src	29	71	¿A úhdityúha tsáma teene áábojcátsi?
dictinterpretation	example-tgt	71	127	¿Tú eres el que provocas los insultos? ‖ acción de...
orthographicinterpretation	headorth	0	13	áá b o j c á ts i
formatting	bold	0	13	áábojcátsi
formatting	italic	14	18	abs.
formatting	italic	29	43	¿A úhdityúha
formatting	italic	44	71	tsáma teene áábojcátsi?
pagelayout	newline	43	43	
pagelayout	tab	44	44	
pagelayout	newline	84	84	

Annotations

```
<?xml version="1.0" encoding="utf-8"?>
<cesDoc version="3.9">


<cesHeader version="2.0" type="text" status="update" date.updated="2010-08-02T12:41:19" date.created="2010-07-13T15:02:34">
    <fileDesc>
        <titleStmt>
            <h.title>Thiesen, Wesley &amp;amp; Thiesen, Eva. 1998. Diccionario Bora–Castellano Castellano–Bora, Entry 5 on Page 27</h.title>
        </titleStmt>
        <editionStmt version="1.0" />
        <publicationStmt>
            <distributor>Research Unit "Quantitative Language Comparison"</distributor>
            <pubAddress>http://www.en.esp.phonetik.uni-muenchen.de/personen/professoren/cysouw/index.html</pubAddress>
            <availability status="free" region="world">http://creativecommons.org/licenses/by/3.0/</availability>
            <pubDate>2010-08-02T12:41:19</pubDate>
        </publicationStmt>
        <sourceDesc>
            <biblStruct>
                <monogr>
                    <h.title>Diccionario Bora–Castellano Castellano–Bora</h.title>
                    <h.author>Thiesen, Wesley &amp;amp; Thiesen, Eva</h.author>
                    <imprint>
                        <pubPlace></pubPlace>
                        <publisher></publisher>
                        <pubDate>1998</pubDate>
                    </imprint>
                </monogr>
            </biblStruct>
        </sourceDesc>
    </fileDesc>
    <profileDesc>
        <langUsage>
            <language indict="src" id="boa" iso639="boa">Bora</language>
            <language indict="tgt" id="spa" iso639="spa">Spanish</language>
        </langUsage>
        <wsdUsage>
          <writingSystem id="utf-8">UTF-8</writingSystem>
        </wsdUsage>
    </profileDesc>
    <revisionDesc>
        <change>
            <changeDate value="2010-08-02T12:41:19" />
            <changeVersion>1.0</changeVersion>
        </change>
    </revisionDesc>
</cesHeader>




<text>
    <body>


<div type="dictentry">
    <p id="27.5">áábojcátsi abs. insultos. ¿A úhdityúha tsáma teene áábojcátsi? ¿Tú eres el que provocas los insultos? | acción de...</p>
    <p id="27.6">[áábójcatsi] vrec. insultarse el uno al otro. ¿Íveeki ámuha máábócatsíhijcyá? ¡Íṁiáámére bo meíjcyaj! ¿Por qué se insultan? ¡Vivan en armonía!</p>
</div>


    </body>
</text>

</cesDoc>
```

```
<?xml version="1.0" encoding="utf-8"?>
<cesAna version="1.5" type="tok" doc="http://www.cidles.eu/quanthistling/source/thiesen1998/27/5/text.xml">


<cesHeader version="2.0" type="text" status="update" date.updated="2010-08-02T12:41:19" date.created="2010-07-13T15:02:34">
    <fileDesc>
        <titleStmt>
            <h.title>Thiesen, Wesley &amp;amp; Thiesen, Eva. 1998. Diccionario Bora-Castellano Castellano-Bora, Entry 5 on Page 27</h.title>
        </titleStmt>
        <editionStmt version="1.0" />
        <publicationStmt>
            <distributor>Research Unit "Quantitative Language Comparison"</distributor>
            <pubAddress>http://www.en.esp.phonetik.uni-muenchen.de/personen/professoren/cysouw/index.html</pubAddress>
            <availability status="free" region="world">http://creativecommons.org/licenses/by/3.0/</availability>
            <pubDate>2010-08-02T12:41:19</pubDate>
        </publicationStmt>
        <sourceDesc>
            <biblStruct>
                <monogr>
                    <h.title>Diccionario Bora-Castellano Castellano-Bora</h.title>
                    <h.author>Thiesen, Wesley &amp;amp; Thiesen, Eva</h.author>
                    <imprint>
                        <pubPlace></pubPlace>
                        <publisher></publisher>
                        <pubDate>1998</pubDate>
                    </imprint>
                </monogr>
            </biblStruct>
        </sourceDesc>
    </fileDesc>
    <profileDesc>
        <langUsage>
            <language indict="src" id="boa" iso639="boa">Bora</language>
            <language indict="tgt" id="spa" iso639="spa">Spanish</language>
        </langUsage>
        <wsdUsage>
          <writingSystem id="utf-8">UTF-8</writingSystem>
        </wsdUsage>
    </profileDesc>
    <revisionDesc>
        <change>
            <changeDate value="2010-08-02T12:41:19" />
            <changeVersion>1.0</changeVersion>
        </change>
    </revisionDesc>
</cesHeader>



<chunkList>


        <chunk from="27.5/0">
            <tok type="dictinterpretation" value="head" from="27.5/0" to="27.5/13">
                <orth>ááb̨ojcátsi</orth>
            </tok>
            <tok type="dictinterpretation" value="pos" from="27.5/14" to="27.5/18">
                <orth>abs.</orth>
            </tok>
            <tok type="dictinterpretation" value="translation" from="27.5/18" to="27.5/27">
                <orth>insultos</orth>
            </tok>
            <tok type="dictinterpretation" value="example-src" from="27.5/29" to="27.5/71">
                <orth>¿A úhdityúha tsáma teene ááb̨ojcátsi?</orth>
            </tok>
            <tok type="dictinterpretation" value="example-tgt" from="27.5/71" to="27.5/127">
                <orth> ¿Tú eres el que provocas los insultos? | acción de...</orth>
```

Displaying **entry 2 on page 1.**

Entry XML
Entry formatting XML
Entry dictionary interpretation XML

ID	Entry	Page
Page 1, ID 2 Link by ID	\lx ambiŋ\ge wing\zn 098\sf animals; birds; animal parts\cf\pg 43\dt 22/Nov/2009	Page: 1, Column 1 View page

Annotations for main entry

Entry text:

\lx ambiŋ\ge wing\zn 098\sf animals; birds; animal parts\cf\pg 43\dt 22/Nov/2009

Python code

Type	Value	Start	End	Substring
dictinterpretation	head	4	10	ambiŋ
dictinterpretation	translation	14	18	wing
pagelayout	newline	10	10	
pagelayout	newline	18	18	
pagelayout	newline	25	25	
pagelayout	newline	57	57	
pagelayout	newline	60	60	
pagelayout	newline	66	66	
pagelayout	newline	81	81	

Top ^

"sparse" annotation of sources, e.g. brat

http://brat.nlplab.org

Summary

- With integration of different resources: **Watch out for data-reduction!**
- Harmonisation will always reduce information, so it should never be “hard-coded”
- Keep all information from the original, and add harmonisation as a separate layer
- ‘head’ and ‘translation’ are already highly interesting!

Graphemic parsing

- **Unicode normalization**
  ő vs. o ˜ ´
- **Orthographic parse**
  separate orthographic units as used in the source: “graphemes”
- **Orthographic harmonization**
  research specific!

Graphemic parsing

- **Code points** (7) t s ʰ o ̃ ́ :
- **Characters** (4) t s^h ő :
- **Graphemes** (2) ts^h ő:

IPA is no solution

- IPA is ‘just another orthography’
- Harmonisation across different IPA transcriptions is still necessary
- Parsing of IPA is not trivial
- Orthography often shows structural information that is difficult to find in the pronunciation (morphology, diachrony)

Translation

- To match meaning across dictionaries, more detailed description of meaning is always better
- Don’t try to reduce the meaning to some pre-established set of ‘core’ meanings
- For translations into major languages, massive computational resources are available!