

Annotations and beyond - what (not) to standardize, and how come that sometimes, it just happens

November 2009, Stockholm, CLARIN-WS



Anton Batliner

Pattern Recognition Lab

Friedrich-Alexander-Universität Erlangen-Nürnberg

Overview



- annotations: some case studies
 - German vowels in SAMPA
 - the arrow
 - how many emotion dimensions
- beyond annotations
 - explicit standards: performance measures
 - implicit standards: tools and type of databases

my personal experience, being a phonetician
who has ended up amongst engineers,
and doing speech research beyond pure ASR

a case study: missing symbols in SAMPA

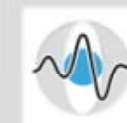


Duden/IPA: bibliote:k

SAMPA: bi: bli: o: te: k / b|bliOte: k

¬SAMPA: bi~~x~~ bli~~x~~ te: k

The official SAMPA page for German



Vowels

The vowels fall into three groups, "checked" (short), "free" (long), and two short vowels that only occur in unstressed position. There is a genuine short-long vowel distinction in German, the long vowels being roughly twice as long (all other things being equal) as the short vowels.

The checked vowels are I E a O U Y ɘ:

I	Sitz	zIts
E	Gesetz	gɘ"zEts
a	Satz	zats
O	Trotz	trOtS
U	Schutz	SUtS
Y	hübsch	hYpS
ɘ	plötzlich	"plɪ9tsɪIC

short/checked/lax

There are 8 pure free vowels, i: e: E: a: o: u: y: ɔ:, and three free diphthongs, aI aU OY:

i:	Lied	li:t
e:	Beet	be:t
E:	spät	SpE:t
a:	Tat	ta:t
o:	rot	ro:t
u:	Blut	blu:t
y:	süß	zy:s
ɔ:	blöd	blɔ:t
aI	Eis	aIs
aU	Haus	haUs
OY	Kreuz	krOYts

long/free/tense

The unstressed "schwa" vowel is:

ɐ	bitte	"bItɐ
---	-------	-------

schwa

SAMPA-D-VMlex, Dokumentation V1.0



Bei Vokalen in Lehnwörtern romanischen oder griechischen Ursprungs besteht oft eine Aussprache­fluktuation, und damit auch eine Variation in der phonemischen Transkription: Ein nicht phonologisch bedingter Wechsel zwischen langen, kurzen, bzw. gespannten und ungespannten Vokalen kommt z.B. in Wörtern wie *Kommunikation*, *Telefon*, *Silikon*, *Dialyse* vor, die zwischen verschiedenen Transkribenten zu Inkonsistenzen führen kann. *Aus theoretischer Sicht wäre die Einführung einer besonderen Vokalklasse für diese Vokale gerechtfertigt; dies wird jedoch hier nicht vorgenommen.*

	Bibliothek	bi:p ljo: 'te:k		FAU
	Bibliothek	bi:bli:ö:te:k	bi:.bli:.o:. +t'e:k	VM
	bibliothek	b bl O'te:k		
	bibliothek	bib li o 'te:k	bib lio 'te:k	DB

1. mapping necessary for ASR
2. rich representation necessary for basic research

another case study



notational conventions

or

Göran Hammarström and the arrow



NP+VP → S

yet another case study



■ emotion dimensions

- two
- three
- many

Emotion Markup Language (EmotionML) 1.0

W3C Working Draft 29 October 2009



2.2.2 The `<dimensions>` element

Annotation	<code><dimensions></code>
Definition	Description of an emotion or a related state using a set of dimensions.
Children	<code><dimensions></code> MUST contain one or more dimension elements. The names of dimension elements which may occur as valid child elements are defined by the <code>set</code> attribute.
Attributes	<ul style="list-style-type: none"> • Required: <ul style="list-style-type: none"> ◦ <code>set</code>, a name or URI identifying the set of dimension names that can be used. • Optional: <ul style="list-style-type: none"> ◦ <code>confidence</code>, the annotator's confidence that the entirety of dimensional annotation given is correct.
Occurrence	A single <code><dimensions></code> MAY occur as a child of <code><emotion></code> .
Annotation	Dimension elements
Definition	Annotation of a single emotion dimension. The tag name must be contained in the list of values identified by the <code>set</code> attribute of the enclosing <code><dimensions></code> element.
Children	Optionally, a dimension MAY have a <code><value></code> child element.
Attributes	<ul style="list-style-type: none"> • Required: <ul style="list-style-type: none"> ◦ (none) • Optional: <ul style="list-style-type: none"> ◦ <code>value</code>, the (constant) scale value of this dimension. ◦ <code>confidence</code>, the annotator's (constant) confidence that the annotation given for this dimension is correct.
Occurrence	Dimension elements occur as children of <code><dimensions></code> . Valid tag names are constrained to the set of dimension names identified in the <code>set</code> attribute of the <code><dimensions></code> parent element. For any given dimension name in the set, zero or one occurrences are allowed within a <code><dimensions></code> element.

Emotion Markup Language (EmotionML) 1.0

W3C Working Draft 29 October 2009



A `<dimensions>` element describes an [emotion](#) or a [related state](#) in terms of a set of [emotion dimensions](#). The names of the emotion dimensions MUST belong to a clearly-identified set of dimension names, which MUST be defined according to [Defining vocabularies for representing emotions](#).

The set of values that can be used as tag names of child elements of the `<dimensions>` element is indicated in the `set` attribute of the `<dimensions>` element. Different sets can be used, depending on the requirements of the use case.

ISSUE: The details of the definition of sets of values need to be sorted out. Throughout this draft, we assume the use of a `set` attribute. Whether a `set` attribute should actually be used, and if so, the format of its attribute values, needs to be clarified in the context of [Defining vocabularies](#). This issue is related to the section [Considerations regarding the validation of EmotionML documents](#).

There are no constraints regarding the order of the dimension child elements within a `<dimensions>` element.

Any given dimension is either unipolar or bipolar; its `value` attribute MUST contain either discrete or continuous [Scale values](#).

ISSUE: the definition of the set of dimensions should include the detailed constraints on valid values of the `value` attribute.

A dimension element MUST either contain a `value` attribute or a `<trace>` child element, corresponding to static and dynamic representations of [Scale values](#), respectively.

If the dimension element has both a `confidence` attribute and a `<trace>` child, the `<trace>` child MUST NOT have a `samples-confidence` attribute. In other words, it is possible to either give a constant confidence on the dimension element or a confidence trace on the `<trace>` element, but not both.

Emotion Markup Language (EmotionML) 1.0

W3C Working Draft 29 October 2009



Examples:

One of the most widespread sets of emotion dimensions used (sometimes by different names) is the combination of valence, arousal and potency. Assuming that arousal and potency are unipolar scales with typical values between 0 and 1, and valence is a bipolar scale with typical values between -1 and 1, the following example is a state of rather low arousal, very positive valence, and high potency -- in other words, a relaxed, positive state with a feeling of being in control of the situation:

```
<emotion>
  <dimensions set="valenceArousalPotency">
    <arousal value="0.3"/><!-- lower-than-average arousal -->
    <valence value="0.9"/><!-- very high positive valence -->
    <potency value="0.8"/><!-- relatively high potency -->
  </dimensions>
</emotion>
```

In some use cases, custom sets of application-specific dimensions will be required. The following example uses a custom set of dimensions, defining a single, bipolar dimension "friendliness".

```
<emotion>
  <dimensions set="myFriendlinessDimension">
    <friendliness value="-0.7"/><!-- a pretty unfriendly person -->
  </dimensions>
</emotion>
```

Different use cases require continuous or discrete [Scale values](#); the following example uses discrete values for a bipolar dimension "valence" and a unipolar dimension "arousal".

```
<emotion>
  <dimensions set="discreteValenceArousal">
    <arousal value="very high"/>
    <valence value="slightly negative"/>
  </dimensions>
</emotion>
```

No fixed standard for content of emotional dimensions



```
<emotion>
  <dimensions set="valenceArousalPotency">
    <arousal value="0.3"/><!-- lower-than-average arousal -->
    <valence value="0.9"/><!-- very high positive valence -->
    <potency value="0.8"/><!-- relatively high potency -->
  </dimensions>
</emotion>
```

some use cases, custom sets of application-specific dimensions will be required. The foll

```
<emotion>
  <dimensions set="myFriendlinessDimension">
    <friendliness value=" 8.7"/><!-- a pretty unfriendly person -->
  </dimensions>
</emotion>
```

and now the alternative: plain ASCII



<word/unit> f07

Classes vs. dimensions in emotion annotation: mapping from early quantization to late and back again



annotation <i>experts/naive, one/more</i>	recognition <i>extraction, selection, ...</i>	decision <i>emotion module</i>	application <i>higher modules</i>
binary n-ary	classification	2 classes n classes	eventually, some binary decision?
continuous	regression	continuous	⇒ <i>to be or not to be</i>

Content vs. form, and mapping



- content difficult (and sometimes fatal) to standardize
- form (notational convention) can and might be standardized – in XML or plain ASCII
- missing (ML-) standardization on the level of annotations seems not necessarily to be a big problem
- mapping (“translation”) between (sub-)standards and levels of representation most important
 - e.g. if you are discontent with ToBI, get rid of everything but the stars

beyond annotations: (good or bad) standards for performance measures (**explicit standards**)



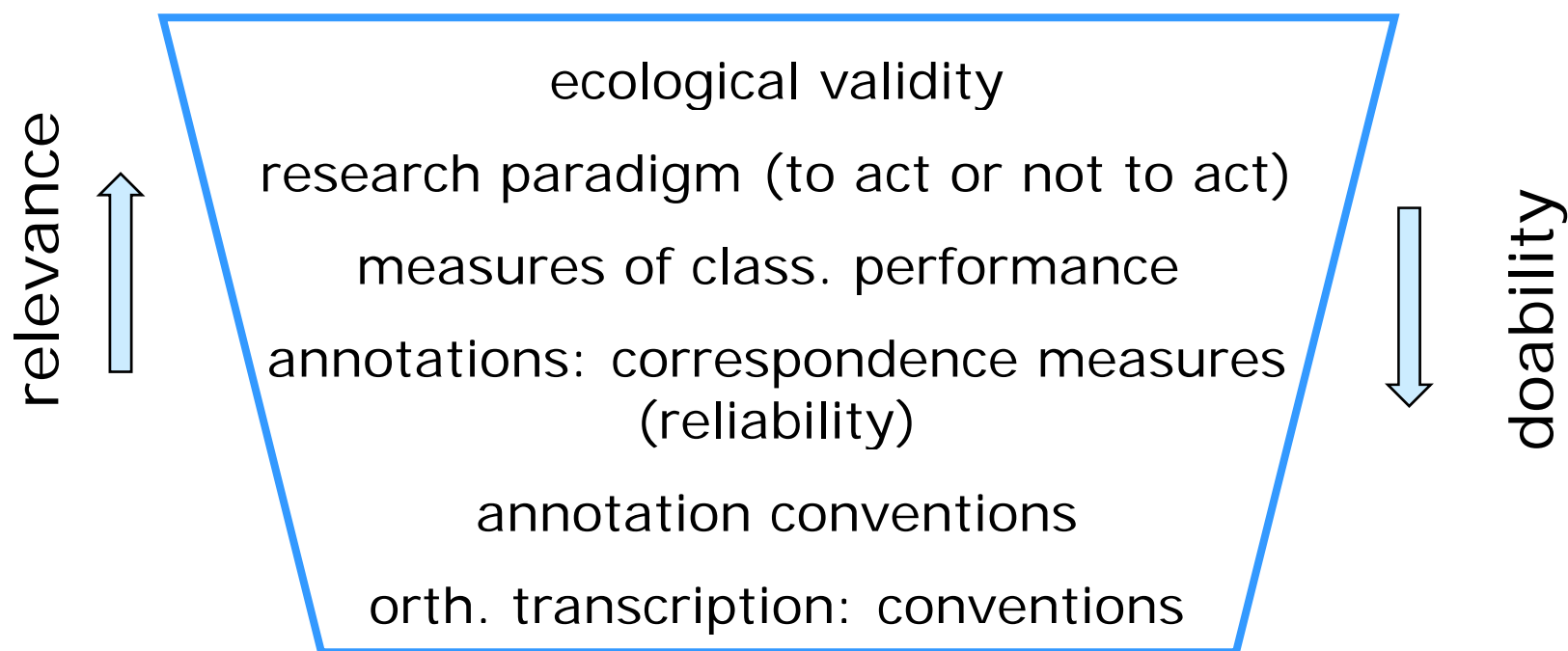
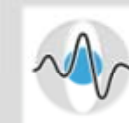
- correspondence measures *weak and controversial*
 - correlations vs. kappa etc.: .2-.4, .4-.6, .6-.8, .8-1.0
poor fair good excellent
- classification measures *just some mess*
 - recall, accuracy, precision, AUR, sensitivity, specificity, F-values, weighted average recall (WA=RR), unweighted average recall (UA=CL), equal error rate, ...
- significance measures *strong and controversial*
 - p-values, alpha: 0.001, 0.01, 0.05

Quasi-/ **implicit standards**, or: being determines consciousness



- WEKA: a big step towards democracy in data mining
 - but:
 - implicit pre-selection of classifiers: no LDC (which is/has been standard in SPSS)
 - but SVM with linear kernel is sort of LDC in disguise
 - selection of sub-sets \Rightarrow 10-fold cross-validation (not speaker-independent)
 - ZeroR (majority class) as baseline “classifier”
- type of data ~ research paradigm
 - to act or not to act?

Levels of standards



what and what not to standardize



- *content* vs. *form* and *mapping*
- pseudo-scientific "*semantisation*" and "*thresholding*" of numbers (kappa, p-values)
- agreement on (the "syntax" of) *evaluation procedures*
- *reliability* or: the story of the needle in a haystack that was searched for in the light of a lantern
- do we want to standardize the bricks in a legoland science or are we aiming at the real stuff (i.e. ecological *validity*)?



Thank you for your attention