Bassey E. Antia

University of Maiduguri, Nigeria
Alexander von Humboldt Fellow, University of Bielefeld, Germany
Corpora and Language Technology for Development: Strategic Management Options for African Higher Education
Introduction

- From using and observing technologies (related to translation & mobile telephony) in the context of Western languages in particular, I have been sensitized to the role corpora play in the conceptualization, development and use of technologies.

- There are at least three conclusions I have drawn from such use & observation:
  - 1. In some environments of the North, corpora constitute one of the drivers of technological advance, a growth that is widening the “digital divide“ (NTIA 2001) – in terms of access to new technologies – and spawning North-South inequalities on a number of indicators, e.g. access to information & knowledge
  - 2. Work on corpora has shown that the popular solution paradigm to the divide – that is, hardware affordability (USD100 laptop; the Brazilian Computador Popular, etc.) & narrowly defined skills – constricts the pool from which change agents/actors can be drawn.
  - 3. Linguistics in African Higher Education has a role and a strategic interest in intervening to address the digital divide.
Aim & structure of talk

1. Advocacy on behalf of corpora and language technology in Africa by:

2. Presenting technologies (related to mobile telephony applications & translation processes automation) in the conceptualization and use of which corpora have played a role.

3. Using (2) as basis to suggest that inattention to corpora in parts of Sub-Saharan Africa potentially hinders the development of, and enjoyment of benefits associated with, corpora-driven technologies in OECD environments.

4. Proposing a strategic management framework within which linguistics in African higher education can think through options to address those aspects of technological inequality that are in part based on corpora.
Corpora & Language Technology: Overview

• Corpora
  - as: collections of free running, naturally occurring texts, stored in electronic form – in which they can be manipulated or exploited.

• Language technology
  - as: tools for processing natural language (synthesis, annotation, retrieval, storage, verification, etc.)

• These are exciting times for corpora in technology
  - Corpora as training basis for machine learning systems (e.g. annotation & retrieval technologies);
  - Paradigm shift in Artificial Intelligence (Expert Systems) from brainstorming to text-based knowledge modeling/ontologies;
  - Markup technologies, and particularly current efforts at creating the Semantic Web (which, through annotation at granular levels, should make information on the Internet more machine-understandable than it currently is);

• Our scope in this presentation:
  - Written corpora in mobile telephony and translation-related applications
Corpora in mobile telephony applications

- Corpora provide basis for:
  - word frequencies (cf. sorting of textonyms)
  - patterns of word combinations (cf. predictive texting)
  - Leveraging previous writing effort (cf. word completion, memory)

- Corpora (particularly marked up) as source for retrieving knowledge in novel ways

- Illustrations:
Corpora in mobile telephony applications [1]

- 26 letters of the English alphabet are distributed over 8 keys.
- Problems:
- Without predictive text, “This is an airport” = 38 presses
  - This = 10 presses
  - is = 7 presses
  - an = 3 presses
  - airport = 15 presses
  - space = 3
  - TOTAL = 38 presses.
- With predictive text, challenge of sorting order
- Illustration: textonyms (3-4-6-3)
Your 74663 is smarter than 968 think.

T9 Translator
Enter a word by pressing just one key per letter to see how T9 works. This simulation has over 30,000 words. Use your phone to get the complete T9 experience.

3463
find

Sending 8398 63772437 shouldn't be about hammering out the ABCs. It should be about 78425 and easy communication.

With T9 Text Input, you can stop 8277464 out the 5388377 and start getting out the 6377243. T9 is predictive text 76389273 that makes it faster and easier to type on small 662453 devices.
Your 74663 is smarter than 968 think.

T9 Translator
Enter a word by pressing just one key per letter to see how T9 works. This simulation has over 30,000 words. Use your phone to get the complete T9 experience.

3463
fine

Sending 8398 63772437 shouldn't be about hammering out the ABCs. It should be about 78425 and easy communication.

With T9 Text Input, you can stop 8277464 out the 5388377 and start getting out the 6377243. T9 is predictive text 76389273 that makes it faster and easier to type on small 662453 devices.
Your 74663 is smarter than 968 think.

T9 Translator
Enter a word by pressing just one key per letter to see how T9 works. This simulation has over 30,000 words. Use your phone to get the complete T9 experience.

3463
dine

发送 8398 63772437 不应该只是关于敲打出ABC。它应该更像 78425 轻松的交流。

With T9 Text Input, you can stop 8277464 out of the 5388377 and start getting out the 6377243. T9 is predictive text 76389273 that makes it faster and easier to type on small 662453 devices.
Your 74663 is smarter than 968 think.

T9 Translator

Enter a word by pressing just one key per letter to see how T9 works. This simulation has over 30,000 words. Use your phone to get the complete T9 experience.

Send 8398 63772437 shouldn't be about hammering out the ABCs. It should be about 78425 and easy communication.

With T9 Text Input, you can stop 8277464 out the 5388377 and start getting out the 6377243. T9 is predictive text 76389273 that makes it faster and easier to type on small 662453 devices.
Why the sequence: find, fine, dine, dime?

Answers from corpus frequencies
About the BNC

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. [more]

Search the Corpus

Type a word or phrase in the search box and press the Return key on your keyboard to see up to 50 random hits from the corpus.

Look up: **find**

You can search for a single word or a phrase, restrict searches by part of speech, search in parts of the corpus only, and much more.

The search result will show the total frequency in the corpus and up to 50 examples. [more information]

News from the BNC
Results of your search

Your query was

find

Here is a random selection of 50 solutions from the 40930 found...

A7L 757 'It took a war to compel the British to look at themselves and find themselves interesting,' Dilys Powell was later to remark.

AM6 760 Now essentially this also relies on the children's using their imaginations to find answers --; as in the case of the ';sewer'; and the ';Lost Valley'; contexts --; but the difference is that (a) each ';find'; is given due weight and attention, is publicly shared, evaluated, accepted or rejected, (b) the focus within which their imaginations can have free range is defined from the beginning, i.e. the puzzle is clear --; ';what could be a logical solution?'; is the name of the game, and (c) the problem-solving does not have to be hurried by action.

AM7 63 Having ';failed'; to find the curriculum or examination version of the Holy Grail for themselves during the sixties and seventies (and having exhausted themselves in the process) the schools are, at the moment, resigned to accepting a string of panaceas from without --; the YTS/TVEI initiatives are now being superseded or subsumed by the National Curriculum cure-all?

APD 1285 Following the formal reconciliation of PLO factions in April 1987, progress was made to achieve better harmony and co-ordination between these groups, and to find a happier balance between the twin goals of national and social justice, a process which fortuitously prepared the committees for the challenge posed by the Uprising.

B1X 3293 You can find out before anyone else who it is that's in trouble over rates or who's in debt or who's emigrating, so that you can get in first and grab their paltry few acres?

B32 362 I went around the place and the units to find out where everything was?

B7J 765 When a rat became a competent maze-runner, Lashley removed part of its cerebral cortex and checked to see whether the animal could still find its way through the maze?

Fertig
Results of your search

Your query was

fine

Here is a random selection of 50 solutions from the 12887 found...

A05 1676 This can be reckoned to contain the succession of anecdotes that occurs and the fine detail of working-class life that is provided.

A16 970 A lower setting gives slight orbital movement for harder materials, while the third setting allows the orbital action to be switched off for a straight reciprocating movement for filing or rasping, or fine cutting in hard materials.

A6W 1201 For fine ride/handling balance, ergonomics, hood design Against Cramped interior, unexciting performance, steering, price?

A70 2502 Pierce small holes on top with a fine skewer.

A00 253 It was an afternoon of high drama and low farce in which Cambridge all but threw away a 15-0 lead before Oxford, with a plethora of errors, nullified two fine tries by their American wing Hein.

A96 1331 But the fact is that organic cheese is fine, but unpasteurised organic cheese, with all the flavour not purged by pasteurisation, is far finer.

A13 1555 Fenway fans have simply grown resigned to seeing fine teams commit one horrendous play to throw away a crucial championship game as good as won.

A1L 51 The only other customers in the `Ravel', that afternoon were a musician with a small Irish fiddle and his companion, a tall skeletal man with the remains of what had once been an extremely fine hat, crammed down over his greasy locks.

A5W 1703 `Fine, just fine, Maclean,' Master Beaton said reassuringly.

BMU 2479 He thought it a fine idea, and presently excused himself --, returning with a bottle of Mrs Westaway's cowslip wine, in which to drink to the new
Results of your search

Your query was
dine

Here is a random selection of 50 solutions from the 242 found...

AOL 1017 Come over and I'll wine you and dine you and make love with you for the rest of your life.?

AHR 1972 And I may dine at journey's end?

AHK 1466 Our choice of venue is usually the Mermaid Restaurant where punters can dine al fresco at white plastic tables, rain or shine, in season or out, risking the wrath of the passing gulls.?

AL7 274 Guests dine sumptuously in the 16th Century oak-panelled, non-smoking restaurant renowned for its award winning food.?

AMO 1517 Discover the old, new and futuristic faces of Acapulco, including the famous cliff divers of la Quebrada, or wine and dine on the world's largest catamaran while taking in the magnificent views of Acapulco by night.?

AN4 3314 I Cal'd at the House, where you said I shou'd dine,?

ANF 253 It was impossible to dine out in view of hungry acquaintances.?

ANR 440 From the age of eight onwards he was allowed to dine with his parents at the Bonaparte family dinners which took place every Monday at the Tuileries, and occasionally at more formal dinners which were held in the Galerie de Diane.?

AR8 1410 Indeed that evening he was once again invited to dine at the Embassy.?

BPII 291 He hoped the artist would not be offended, but as he had dined there many years, if it would be a convenience during his present work to dine until it was
Results of your search

Your query was
dime

Only 39 solutions found for this query

A6C 447 My father was one of them, and when "Buddy Can You Spare a Dime?", topped what was then the hit parade, its lyrics summed up the poverty, dependence and despair of the millions of poor throughout the world.

A6C 448 We all knew how little and how vital a dime was.

ACP 756 He's a man who made a lot of money and never had a dime on him.

AJN 69 It cost me two years of impoverishment, spending every dime on legal fees.

ALD 199 "One thing we don't need in this subcontinent," Professor Rokeya Kabeer observed, "is spiritual guidance --; we have gurus a dime a dozen."

ALH 1341 The old man was begging him for a dime to buy a cup of coffee.

ASV 1733 He'd run up another massive telephone bill and hadn't paid Betty a dime.

ATE 2339 "Well, you can still sing for me," said Bernie, "if I give you a dime."

ATE 2347 Whether it was the incentive of the dime or the desire to get it over, there was little hesitation this time, and her hand shot out imperiously at the end of it.

ATE 2402 All she did was hold out her hand and say, Look, Ma, Bernie gave me another dime.

BIL 633 "JINKY CAN YOU SPARE A DIME."?
Intelligent Applications

To power Nokia's natural language technology, MIT's Katz is using a software system he developed in 1993 called Start, which interprets human questions and finds answers using websites.

Here's how the Web version of Start works: users type a question into a text field. The software interprets the query, decides where to seek the answer (in its database or on another website), and responds with a written explanation, a link to a website, or an image.

The Start system understands English sentences by breaking them down into a series of relationships between object, property, and value. For instance, if one types, "What is the population of Iraq?", Start interprets the query: the object is Iraq, the property is population, and the value is what Start seeks.

http://www.technologyreview.com/Infotech/16745/?a=f
Summary: Corpora in mobile technology applications

- Corpora driving innovative technologies that:
  - exploit frequencies
  - retrieve knowledge from text repositories

- ... with implications for
  - entrepreneurship
  - enhanced access to information
  - relevance of linguistics scholarship to the digital divide
  - high return on consumer investment in gadgets

Without exploitable corpora, these technologies may be unavailable in African languages, or only available after delays and at a higher development cost to manufacturers – who might have preferred to simply buy licenses to use such corpora, rather than have to create them.
Corpora and translation processes automation

Case studies of:
1. Translation Memory
2. Text alignment
3. Bilingual term extraction
Corpora in translation-related processes automation

1. Corpora as evidence of previously unrecognized or unleveraged repetition in language:

   “The creative potential of language is undeniable, but the concordances to a corpus remind us forcibly that in most of our utterances we are creators of habit, immensely predictable, rehearsing the same old platitude and same old clichés in almost everything we say. If it were not so, language would become unworkable. Humankind cannot bear very much creativity”. (Hanks 1996)

2. Corpora as basis for posing and answering questions, which then provide heuristics for processing language

   Across texts of different languages, to what use can we put corpus-supported answers to questions concerning (1) frequency of co-occurrence of words in sentences or sentence groups; (2) volume of words/characters; (3) the co-occurrence of digits, proper names, and cognates?

1 & 2 make corpora the basis for Translation Memory, Bilingual Text Alignment (Techniques: length, lexical, geometrical, K-vec, etc.), Bilingual Term Extraction

Illustrations:
Translation Memory

• Stores source-language text segments and their translations, then makes these translations available for reuse.

• Vision underlying development: You never have to translate the same sentence twice.

• Illustration: Leveraging previous translations with SDL Trados Workbench
1. A database in SDL Trados of user's previous German to English translations on Nokia batteries

2. A new German text on Nokia batteries meant for translating into English
1. Opens SDL translation memory to insert any matches found.

2. For the German Word.doc phrase to be translated, there exists in the Memory an exact match with an English translation.

3. The pre-existing English translation is automatically inserted.
Text Alignment

- (Semi-automated) process of connecting parts (words, sentences, paragraphs) of a source-language text to corresponding target language segments.

- Purposes:
  - populating database of Translation Memory;
  - extracting bilingual terminology for human/machine use;
  - development of machine translation systems (statistical, example-based)
  - ...

- Illustration: SDL Trados WinAlign
The strategy: from farm to fork

Food safety is vitally important. People in Europe want to be sure that, wherever it comes from, and wherever they buy it, their food is safe and wholesome. That means setting Europe-wide standards and taking Europe-wide action to enforce them. Doing so is a priority for the European Union.

Work to improve food safety is going on all the time, but there has in addition been a major overhaul in the last couple of years. This was a response to headline-grabbing food safety scares.

Die Strategie: vom Erzeuger bis zum Verbraucher

Die europäischen Verbraucher möchten, dass ihre Lebensmittel sicher und gesund sind. Die EU bemüht sich, dafür zu sorgen, dass die Lebensmittel, die wir verzehren, für alle Bürger den gleichen hohen Standards genügen, ob diese Lebensmittel nun aus dem In- oder Ausland, aus der EU oder aus anderen Ländern stammen.

Es wird standig daran gearbeitet, die Lebensmittelsicherheit zu verbessern, darüber hinaus wurde das Lebensmittelrecht in den letzten Jahren in umfassender Weise auf den neuesten Stand gebracht. Dies geschah als Reaktion auf die schlagzeulastischen
The strategy: from farm to fork

Food safety is vitally important.

People in Europe want to be sure that, wherever it comes from, and wherever they buy it, their food is safe and wholesome.

That means setting Europe-wide standards and taking Europe-wide action to enforce them. Doing so is a priority for the European Union.

Work to improve food safety is going on all the time, but there has in addition been a major overhaul in the last couple of years.

This was a response to headline-hitting food safety scares in the 1990s about such things as 'mad cow' disease, dioxin-contaminated feed and adulterated olive oil.

Die Strategie: vom Erzeuger bis zum Verbraucher

Die europäischen Verbraucher möchten, dass ihre Lebensmittel sicher und gesund sind.

Die EU bemüht sich, dafür zu sorgen, dass die Lebensmittel, die wir verzehren, für alle Bürger den gleichen hohen Standards genügen, ob diese Lebensmittel nun aus dem In- oder Ausland, aus der EU oder aus anderen Ländern stammen.

Es wird ständig daran gearbeitet, die Lebensmittelsicherheit zu verbessern, darüber hinaus wurde das Lebensmittelrecht in den letzten Jahren in umfassender Weise auf den neuesten Stand gebracht.

Dies geschah als Reaktion auf die schlagzeilenstärksten Lebensmittelskandale in den 90er Jahren im Zusammenhang mit BSE, dioxinverseuchten Lebensmitteln und gepanzertem Olivenöl.

Die EU-Rechtsvorschriften zur Lebensmittelsicherheit sollten nicht nur dem neuesten Stand der Erkenntnisse entsprechen, dem Verbraucher sollten auch so viele Informationen wie möglich darüber an die Hand gegeben werden.
From Alignment to Term Extraction

- Extraction as process of identifying candidate terms from texts

- Purposes:
  - Populating term bases/publishing dictionaries
  - Populating machine-readable lexica
  - Knowledge engineering processes (e.g. ontologies in support of translation)

- Illustration: TermCalc (developer: Tom Vanallemeersch)
TermCalc for candidate terms in EU food safety texts

Human validated items of interest

<table>
<thead>
<tr>
<th>Score</th>
<th>SentLinked</th>
<th>SentFreq1</th>
<th>WordGroup1</th>
<th>WordGroup2</th>
<th>Validated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.93</td>
<td>72</td>
<td>74</td>
<td>eu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.93</td>
<td>6</td>
<td>7</td>
<td>label</td>
<td>etikett</td>
<td></td>
</tr>
<tr>
<td>0.92</td>
<td>5</td>
<td>5</td>
<td>cap</td>
<td>gap</td>
<td></td>
</tr>
<tr>
<td>0.91</td>
<td>9</td>
<td>9</td>
<td>EFSA</td>
<td>behörde</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>12</td>
<td>12</td>
<td>commission</td>
<td>kommission</td>
<td></td>
</tr>
<tr>
<td>0.88</td>
<td>6</td>
<td>8</td>
<td>example</td>
<td>beispielsweise</td>
<td></td>
</tr>
<tr>
<td>0.87</td>
<td>21</td>
<td>27</td>
<td>rules</td>
<td>vorschriften</td>
<td></td>
</tr>
<tr>
<td>0.86</td>
<td>8</td>
<td>11</td>
<td>information</td>
<td>informationen</td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>11</td>
<td>13</td>
<td>Farmers</td>
<td>landwirte</td>
<td></td>
</tr>
<tr>
<td>0.83</td>
<td>5</td>
<td>6</td>
<td>safe food</td>
<td>sichere</td>
<td></td>
</tr>
<tr>
<td>0.82</td>
<td>21</td>
<td>29</td>
<td>feed</td>
<td>lebens</td>
<td></td>
</tr>
<tr>
<td>0.82</td>
<td>7</td>
<td>7</td>
<td>genetically modified</td>
<td>gentechnisch</td>
<td></td>
</tr>
<tr>
<td>0.81</td>
<td>5</td>
<td>8</td>
<td>new</td>
<td>neue</td>
<td></td>
</tr>
<tr>
<td>0.81</td>
<td>5</td>
<td>8</td>
<td>scientific advice</td>
<td>gutachten</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>6</td>
<td>8</td>
<td>want</td>
<td>wollen</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>27</td>
<td>36</td>
<td>food safety</td>
<td>lebensmittelsicherheit</td>
<td></td>
</tr>
<tr>
<td>0.77</td>
<td>5</td>
<td>7</td>
<td>ingredients</td>
<td>zutaten</td>
<td></td>
</tr>
<tr>
<td>0.77</td>
<td>5</td>
<td>7</td>
<td>animal health</td>
<td>tiergesundheit</td>
<td></td>
</tr>
</tbody>
</table>

§4. **Animal health** and welfare

§4. **Tiergesundheit** und artgerechte Tierhaltung

119: The European Commission consults the Standing Committee on the Food Chain and **Animal Health**, on which all EU countries are represented.

Summary: Translation processes automation

• Technologies & techniques conceptualized on the basis of knowledge in corpora, and these technologies are then deployed to the leveraging of parallel corpora:

  – for populating databases (e.g. Translation Memory, but also Terminology) which in turn makes it possible to …

  – achieve speed, consistency & cost-savings in translations done in domain-specific contexts
Corpora-technology interface

Some African issues
Overview

- In spite of islands of activity…
  - U.S.-based activities on Machine Translation (University of Maryland Eastern Shore, Albion College, Michigan …)
  - Finland-based Kishwali Machine Translation
  - South Africa's human language technology programme
  - Morocco's Arabic language technology projects
  - German (Bielefeld)-Nigerian (Uyo) Ibibio language synthesis projects (speech)
  - Etc.

- a picture of relative inaction, probably explained in part by following hypotheses:…
African language presence on the Web

Table 1: Didi-Kidiri & Baboya (2003)

<table>
<thead>
<tr>
<th>Description of status/feature of language</th>
<th>Samples of words, expressions, or whole lexica</th>
<th>Texts, typically (translations of) legal, admin, religious, literary documents</th>
<th>Sites for language learning, typically with English, French, German, Spanish, being the media</th>
<th>African language as web interface medium in a sub-sample of 503 sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nothing other than a mere mention of name of language</td>
<td>76 (14.4%)</td>
<td>76 (7.4%)</td>
<td>89 (8.7%)</td>
<td>33 (6.56%)</td>
</tr>
<tr>
<td>Listing of bibliographic resources on language</td>
<td>383 (27.9%)</td>
<td>152 (14.8%)</td>
<td>24 (2.33%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Van der Veken & de Schryver (2003)

<table>
<thead>
<tr>
<th>Language</th>
<th>Word types</th>
<th>Word tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausa</td>
<td>30 996</td>
<td>858 734</td>
</tr>
<tr>
<td>Somali</td>
<td>40 251</td>
<td>304 361</td>
</tr>
<tr>
<td>Lingala</td>
<td>11 557</td>
<td>193 772</td>
</tr>
<tr>
<td>IsiXhosa</td>
<td>149 553</td>
<td>943 772</td>
</tr>
</tbody>
</table>
1. African language presence on the Web

- A device for retrieving information on the Web (to empower users) is only as useful as the number of knowledge-rich texts available on the Web in the user’s particular language, except of course a translation module is added to the device, as in the Indian Simputer.

- Besides, without large bodies of local language texts, there is unlikely to be enough input for technologies that operate with statistical probabilities and/or require training corpora (predictive texts, machine translation) to be implemented or to generate satisfactory results.
2. Attitude to written use of African languages

- Admittedly, concerns of upward mobility may continue to justify the use of non-African languages in many professional domains;

- yet, written corpora for technological exploitation is not all about recondite and rocket science exposés.

- While the latter are important for reasons of corpus balance, the text requirements of language technology applications can already be quite well served by non-abstruse explorations in writing of lived, daily reality that have the effect of increasing word token levels beyond what is suggested in Table 2.
3. Text management practices

- Many print and electronic mass media generate bi- or multilingual text material in the form of news items translated from a Western language into national languages.

- These multilingual resources are seldom leveraged for in-house needs, let alone processed and made available for other, external uses.

- Some research findings (Antia 1992; 1996):
  - no documentation requirement for the texts of translated news material;
  - physical copy of a translator’s rendering of a news bulletin is discarded, irrespective of how much of the material will be relevant for a subsequent news bulletin that falls within a colleague’s work shift;
  - for reasons of funds, recordings of translations on tape are routinely wiped out so that tapes can be reused;
  - sight translation tends to be viewed as the hallmark of consummate competence.

- Little wonder that unleveraging of previous translations leads to:
  - Translator complaints of insufficient time to produce their texts
  - Truncation of news material + Terminological inconsistency (e.g. kosuwa zar tigbe yezoma; kosuwa kawudataye.
  - Dearth of funds for institutional development
  - Dearth of exploitable multilingual corpora for lexicography (à la Bahlburg & McIntyre 1991), for training machine (learning) systems…
4. Institutionalized paradigms in academic linguistics research

- Substantial research of a theoretical nature in linguistics driven by a rationalist or mentalist epistemology that promotes innate, internalised knowledge over empiricism in the modeling of language.

- On this view, corpora, as an instance of performance data, considered poor reflections of ideal competence because of their many non-linguistically relevant entanglements.

- Applications of this research programme to descriptions of the structure of particular languages have proceeded on the basis of innate knowledge or of respondents’ insights, thus without this method being rationalized in terms of the expediency of collecting or using naturally occurring evidence.

- Even with improving hardware access, relative dearth of active research programmes that emphasize text production and corpus management, or that conduct statistical analyses of occurrences of language features, or that set premium on text analytical studies.
Bridging the digital divide through corpora-technology interface

Strategic Management Options for African Higher Education
A Model of Strategic Planning

Vision → Strategic analysis → Strategic plan → Operative instruments → Strategy control

Feedback loop

Cf. Ziegele (2007)
1. Vision options

- A range of vision thrusts are conceivable. Vision of commitment to corpora and language technology could seek to:
  - empower new citizenry demographics,
  - have language technology products and services contribute to gross domestic product or institutional revenue base by specified margins,
  - enhance institutional and national standing.
## 2. Strategic analysis [1]: Strengths & Opportunities

<table>
<thead>
<tr>
<th>Positive</th>
<th>Internal environment</th>
<th>External environment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>What are the strengths warranting vision or supportive of implementation?</td>
<td>What are the opportunities warranting vision or supportive of implementation?</td>
</tr>
<tr>
<td></td>
<td>Expertise in languages &amp; increasingly (but uncoordinated) in computing</td>
<td>Huge African consumer market for technologies with linguistic knowledge</td>
</tr>
<tr>
<td></td>
<td>Yearly graduating student research projects</td>
<td>International market for resources in African language</td>
</tr>
<tr>
<td></td>
<td>Ecologically-sensitive mandate of universities</td>
<td>Existing international networks on language &amp; technology interface</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Available funding for international cooperation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Global competition in higher education sector</td>
</tr>
</tbody>
</table>
## 2. Strategic analysis [2]: Weaknesses & Threats

<table>
<thead>
<tr>
<th>Negative</th>
<th>Internal environment</th>
<th>External environment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>What are the weaknesses warranting vision or anticipated in implementation?</strong></td>
<td><strong>What are the threats warranting vision or anticipated in implementation?</strong></td>
</tr>
<tr>
<td></td>
<td>Manpower with a convergence of interests in language and computing</td>
<td>Rationalisation of poor performing study programmes</td>
</tr>
<tr>
<td></td>
<td>Phobia for technology among an older generation of relevant actors who may be occupying decision-making positions relevant to vision</td>
<td>Uncertainties of the socio-political environment; preference for other world regions in allocation of international funds</td>
</tr>
<tr>
<td></td>
<td>Feeling of exclusion among linguists vis-à-vis emerging plethora of discourse communities applying technology to language; constricted access to academic opportunities (e.g. publishing, collaboration, etc.)</td>
<td>Requirements of national-level counterpart funding in possible financing arrangements</td>
</tr>
<tr>
<td></td>
<td>Insolvency of language graduates in the (professional language) labour market</td>
<td>Difficulty of finding partners</td>
</tr>
<tr>
<td></td>
<td>Poor infrastructure (electricity, bandwidth, etc.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low funding levels</td>
<td></td>
</tr>
</tbody>
</table>
3. Strategic plan: 3Ms of **Manpower**, Money & Material

<table>
<thead>
<tr>
<th>Options</th>
<th>Elaboration</th>
</tr>
</thead>
</table>
| Training own staff       | 1. MAs/PhDs in facets of language-technology interface  
2. Sandwich programmes or short-term training (Summer Schools)  
3. Attendance at conferences                                                                                                                                   |
| Leveraging available expertise and means | 1. Curriculum innovation: select courses in linguistics and computing for students of both disciplines  
2. Dedicating a fraction of graduating student research to corpora and language technology themes  
3. Joint seminars for staff in linguistics and computing on possible collaborative project themes  
4. Language instruction to enable staff and students in the university community write local language summaries of research they produce  
5. Basic IT skills (e.g. HTML, Web publishing) to enable staff and students post HTML documents |
| Recruitment              | 1. Head-hunting for (local and international) talent in language technology to join faculty on full-time, part-time or visiting basis  
2. Attractive conditions, such as are currently enjoyed in Nigeria by medical and veterinary academics                                                                                           |
| Collaboration            | 1. Identification of regional and international teams for cooperation on projects of common interest; internships, etc.                                                                                         |
|                          | 2. Identification of Diaspora and alumni expertise for capacity-building                                                                                                                                       |
### 3. Strategic plan: 3Ms of **Manpower, Money & Material**

<table>
<thead>
<tr>
<th>Money</th>
<th>Options</th>
<th>Elaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Local, institutional funds</strong></td>
<td>1. Faculty, Interuniversity research funds</td>
<td>2. Trade in language resources (in cash or by barter)</td>
</tr>
<tr>
<td></td>
<td>3. Consultancy</td>
<td></td>
</tr>
<tr>
<td><strong>National advocacy for:</strong></td>
<td>1. language technology as national strategic interest issue, with infrastructural and other budgetary requirements</td>
<td>2. eliciting national private sector funding</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>International sources</strong></td>
<td>1. Competitive research funds</td>
<td>2. Student scholarships and other funded staff exchange programmes</td>
</tr>
</tbody>
</table>
3. Strategic plan: 3Ms of **Manpower, Money & Material**

<table>
<thead>
<tr>
<th>Materials</th>
<th>Options</th>
<th>Elaboration</th>
</tr>
</thead>
</table>
| **Donations** | 1. Through equipment donation schemes of scholarship programmes  
2. From institutions re-kitting  
3. From manufacturers, with or without conditions like offer to localise software or documentation, to participate in testing, etc. | |
| **Sharing** | 1. In the context of project partnerships | |
| **Purchase** | | |
Conclusion

- Measures such as outlined should see African language corpora evolving and serving as basis for:
  - activities that can accelerate the bridging of the digital divide
  - broadening the pool of actors involved in bridging activities to include linguists
  - furthering the strategic interests/mandate of higher education
Thank you for your attention!

Bassey E. Antia

URL: www.basseyantia.net

Email: bantia1@yahoo.co.uk

- This presentation was prepared in the context of a research stay in Germany financed by the Alexander von Humboldt Foundation. The Foundation is thanked for its support.