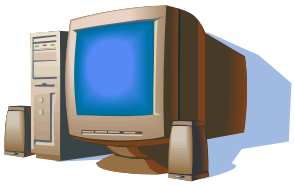
Diacritization: A Challenge to Arabic Treebank Annotation and Parsing

College of Computer Science and Engineering
Information and Computer Science Department

ICS482 – NLP/ Presentation
Term (062)

King Fahd University of Petroleum and Minerals



Presents

Prepared for
Dr. Husni Al-Muhtaseb

By
Al-Elaiwi Moh'd

ID #: 223160

May 13, 2007

- A summary of your lecture

Arabic diacritization (referred to sometimes as vocalization or vowelling), defined as the full or partial representation of short vowels, *shadda* (consonantal length or germination), tanween (*nunation* or definiteness), and *hamza* (the glottal stop and its support letters), is still largely understudied in the current NLP literature. In this lecture, the lack of diacritics in standard Arabic texts is presented as a major challenge to most Arabic natural language processing tasks, including parsing.

Recent studies about the place and impact of diacritization in text-based NLP research are presented along with an analysis of the weight of the missing diacritics on Treebank morphological and syntactic analyses and the impact on parser development.

This is the table of contents:

- List of references

Diacritization: A Challenge to Arabic Treebank Annotation and Parsing
By Mohamed Maamouri, Ann Bies, Seth Kulick Linguistic Data Consortium, University of Pennsylvania, USA

- Obstacles you have faced though the process of preparing and presenting your lecture

1. **I have faced some problem choosing this paper to give a presentation about because there were a lot of choices and I was not sure which one is better to present that have information related to our course topics.**
2. **Reading and abstracting the information that will be easy for the student to understand.**
3. **There were some words that I had to get there meaning from the dictionary.**
4. **My paper was long and I had problem to decide which part to leave and which part to discuses.**

- Things you have learned and skills you have practiced

1. **I have learned how to search and choose a topic that is related to our course.**
2. **I have learned how to summaries and abstract information from a research paper.**
3. **I have learned the good presentation skills.**
4. **I have learned how to manage my time during a limited presentation period.**
5. **I have learned that you should practice a lot before you present your topic.**

- Recommendation

It is a good opportunity to explore our research side and to prepare for the future for the working environment.

I hope that the instructor also the student mention the good and the bad things the presenter have done after he finish the presentation to improve his skills.

- Three true/ false questions addressing the main issues of your lecture with their answers for possible inclusion in the exam.

1. **Is Diacritization a Challenge to Arabic Treebank Annotation and Parsing?**

Yes, it is. Because in Arabic there are a lot of words written the same way but has different meanings. This is way we need Diacritization to distinguish between them.

2. **Is Diacritization Annotation and Parsing applicable for all languages?**

No, it is not. It is only applicable for Arabic language because of its complex linguistic structure and the specific features of its orthographic system.

3. **Is It Important to Use the Diacritization Parsing?**

Yes, it is because it is the way to differentiate between the words that are written similarly.