

Irmi Salminger
Stephan Lücke

Strutture di subordinazione in calabrese

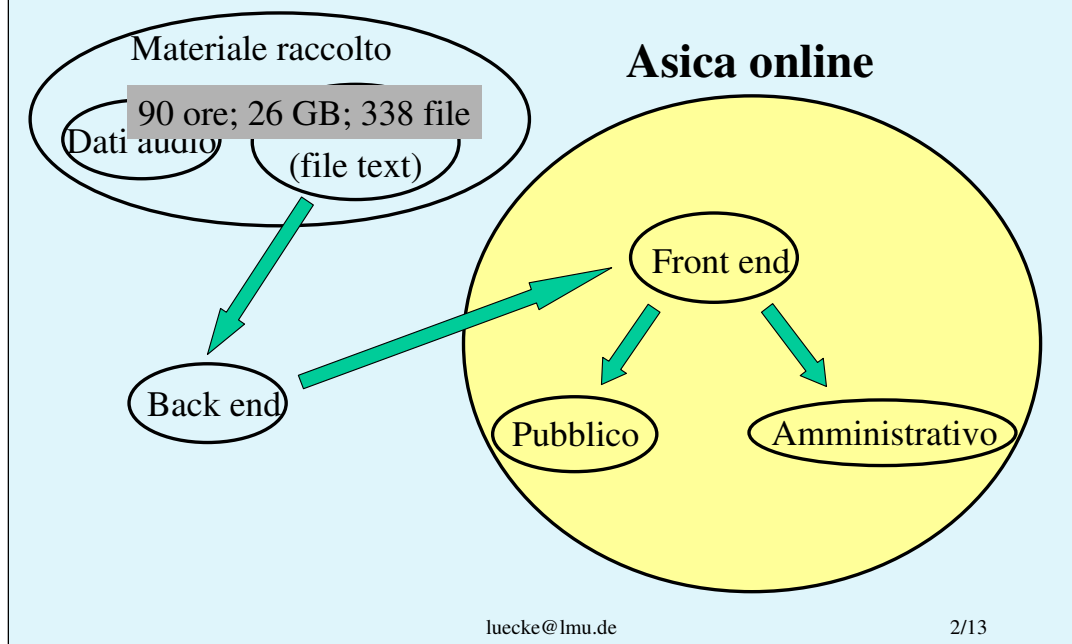
Parte I (di Stephan Lücke):
Aspetti informatici

Ladies and Gentlemen!

First I would like to thank you for your invitation and for the opportunity to present our „Asica“ project.

I am aware that some of you might be offended by the fact that I hold this presentation in English and would like to thank you in advance for your generosity in tolerating it.

My aim is in some way hazardous: to explain a complex technical matter in an understandable manner within ten minutes. I beg your understanding if therefore I concentrate on some aspects I regard as crucial.



My first slide gives an overview of the Asica-Project as a whole. As you can see „Asica“ consists of several members:

The data-core is formed by the material that was collected mainly by Mrs Salminger during 2004 and 2005. It consists of nearly 90 hours of spoken speech stored in 338 audio-files comprising around 26 Gigabytes of data. The major part of these audio-files has been transcribed and is thus available in the form of computer-readable textfiles.

The transcribed material has been imported into a MySQL-database forming the backend of Asica-online which on its part represents the frontend of the system.

- Easy and widespread access
- Low cost
- Fast and easy implementation
- Easy administration
- Platform independence (Windows – Unix – Mac OS)

I will now focus on the technical aspects of our project.

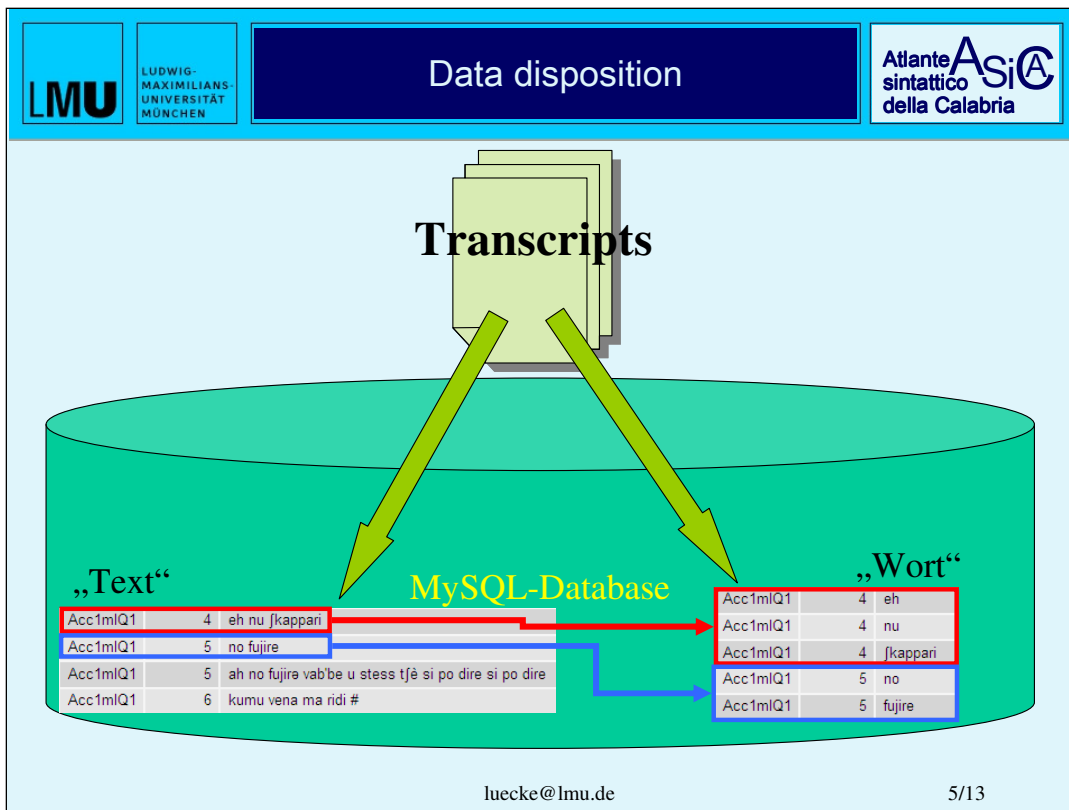
The issue to visualise specific morphosyntactic structures to be found in the collected material in form of geographical maps was basically connected with the following top level requirements:

- The atlas should be easily accessible and from everywhere
- Implementation as well as administration must not involve high costs
- The whole system had to be implemented quickly and simply
- Once running, the system should be easy to administer
- And finally the system should work with any operating system – be it Windows, Unix or Mac

- Internet based client-server architecture
- Open source software:
 - HTML
 - PHP
 - Javascript
 - MySQL

To meet as many of these demands as possible we decided to use an internet-based client-server architecture.

The whole system uses only open source software which is free of charge, transparent and – more or less well – documented. For those of you who are familiar with these things I mention the terms HTML, PHP, Javascript and MySQL.

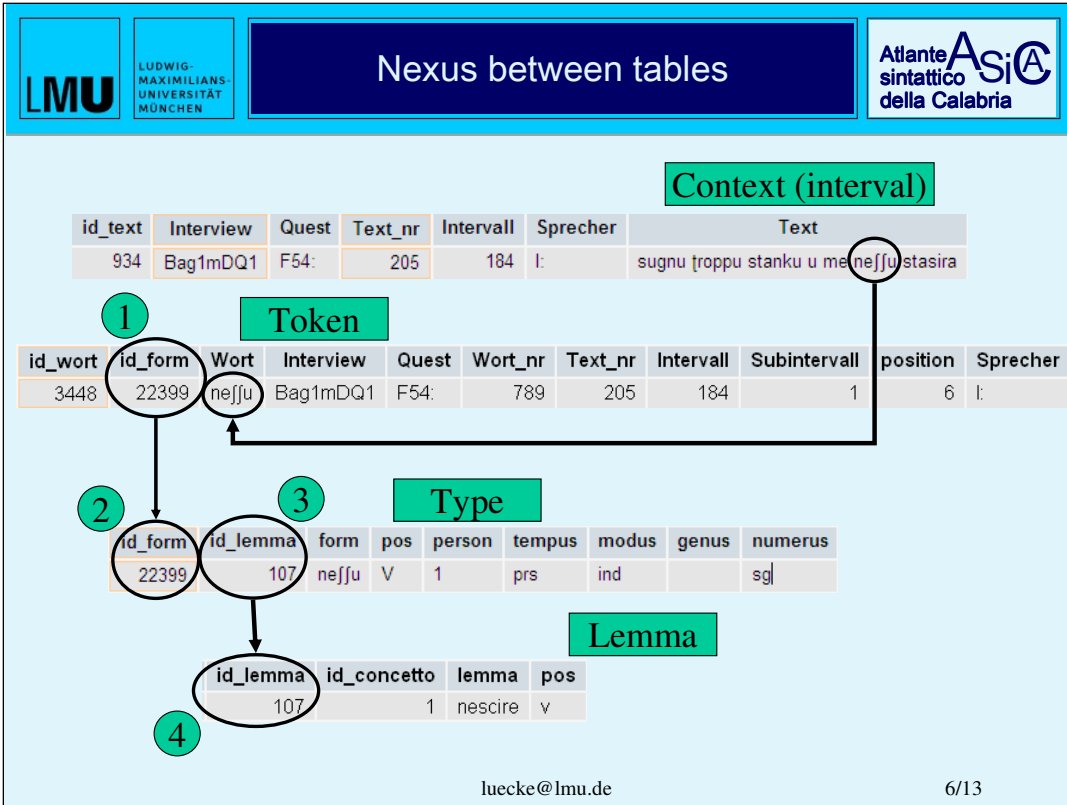


The choice of a MySQL-database asked for a specific treatment of the written transcripts Mrs Salminger has produced. By default any – so called – „relational“ database (such as MySQL) requires data to be entered in tabular form. The precise design of these tables depends on various factors and is at last up to the one in charge.

We decided to duplicate the whole material and to deposit it in two separate tables. In one of these tables which we called „text“ – this is the name that actually is used in the database – each line contains exactly one „interval“. These intervals were defined by Mrs Salminger during the creation of the transcripts using the program Praat. The definition was carried out intuitively, the only purpose being to produce small parts that could later easily be referred to.

The other table containing the material is called „wort“. Here each line contains exactly one word of the transcript. The correlation of the two tables is provided by the fact that certain columns of the table „wort“ contain the corresponding name of the interview and the number of the interval.

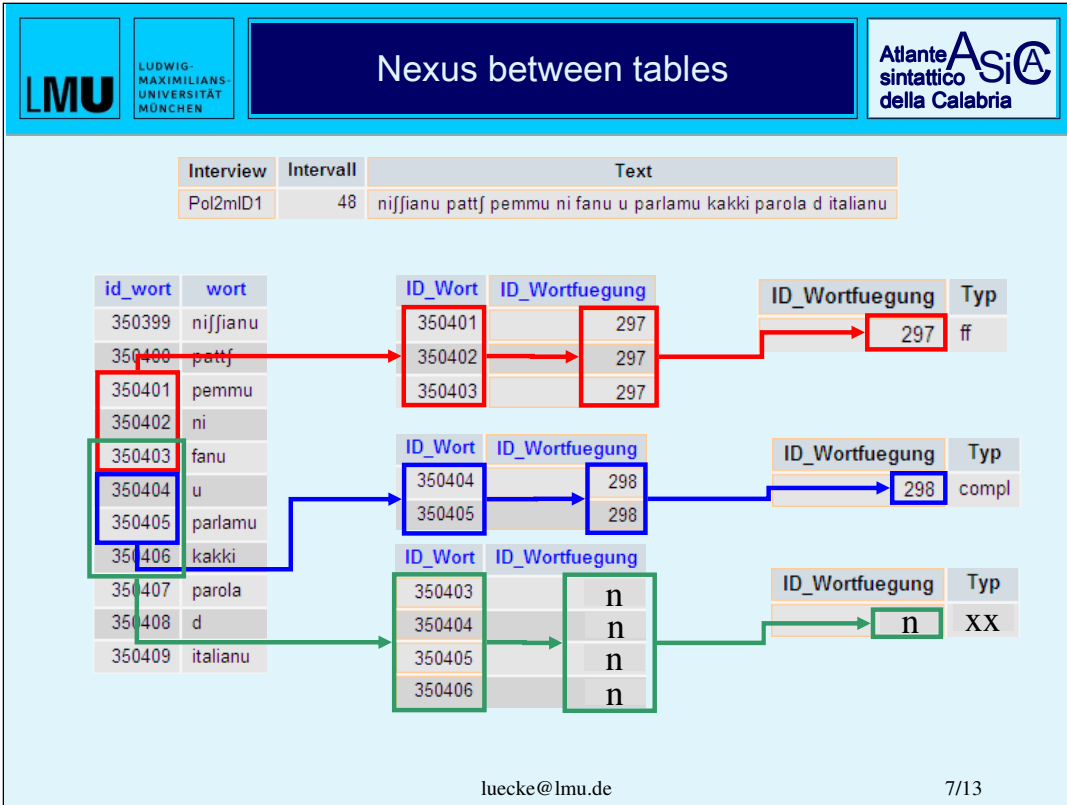
[[[The duplication of the material and its organisation in two different tables has advantages and – as has turned out later – also some disadvantages.]]]



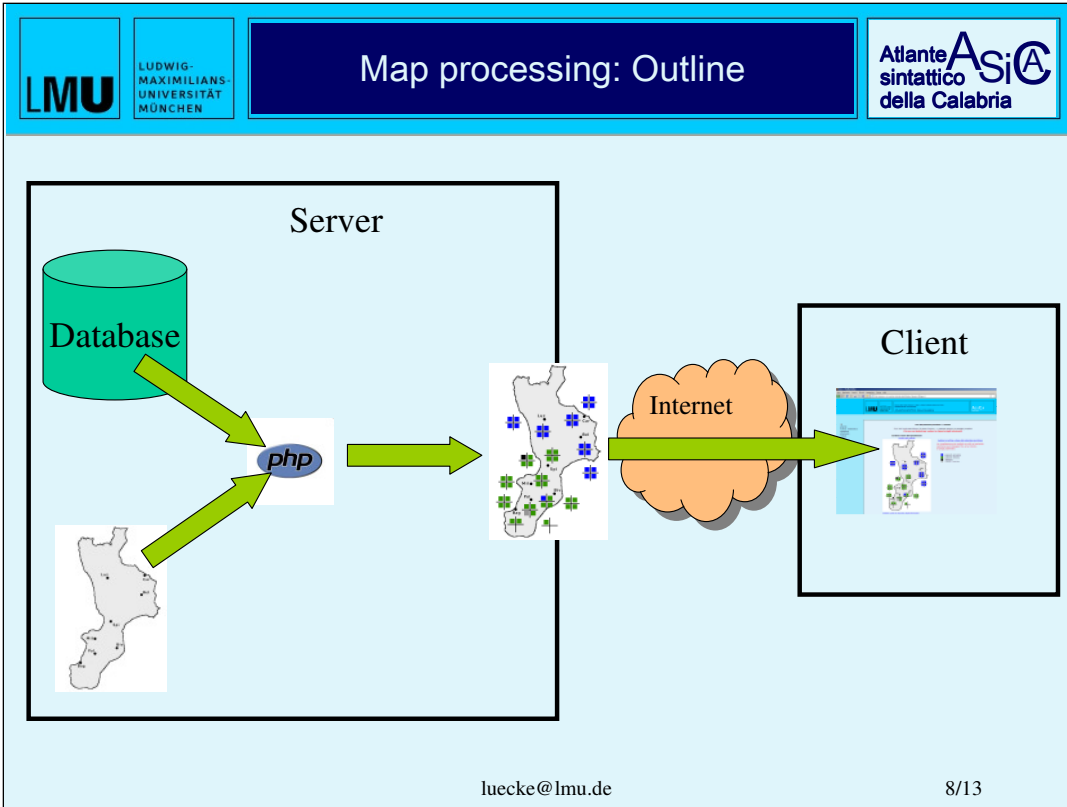
One of the main issues then was the tagging of the material: Round about 350 000 tokens in the table „wort“ had to be identified regarding the main morphological categories such as „part of speech“ (POS), mode, number, grammatical gender and so on. On this task we searched for strategies to reduce data volume as well as effort to a minimum. Therefore we decided not to record the tags in the table „wort“ connected directly to each token but to generate a third table containing the so called „types“ representing unique entities regarding their morphological identification. The tagging now is carried out by attributing each token to its corresponding type. This procedure also minimizes the risk of errors as the morphological tags only have to be fixed once for each type.

Finally a fourth table contains the headwords („lemmata“) to which the types are associated in just the same way the tokens are associated with the types.

The sketched system allows for example to search for any token that is connected with the word „nescire“. Other possible queries would be to look for the various morphological types that are linked with a certain headword or to count and to compare the frequency of the use of a specific form, for example the *passato remoto*, in relation to one or more speakers or groups of speakers. The possibilities are almost boundless.



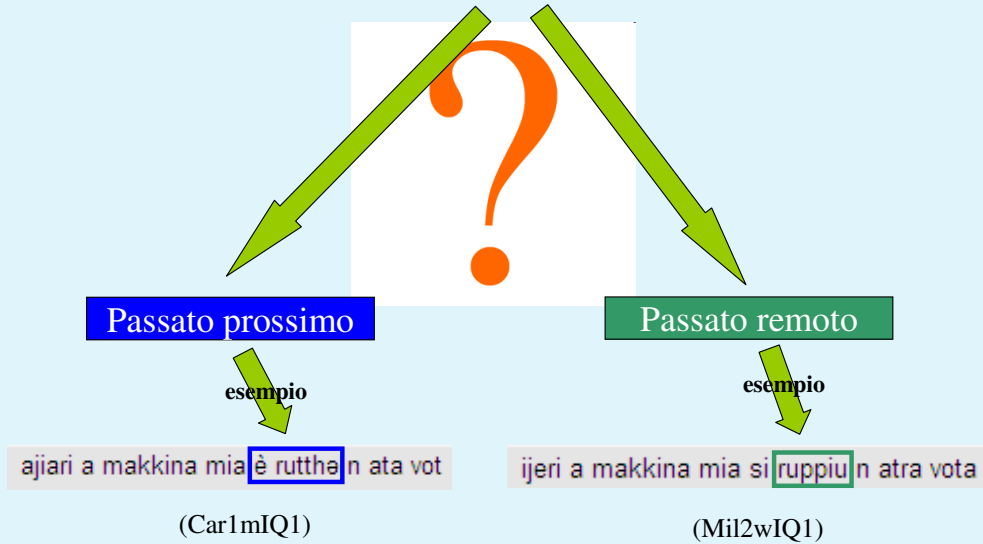
As Mrs Salminger will explain in her contribution some morphosyntactical phenomena can not be found unfaillingly if search is restricted to the morphological level only. In such cases our system offers the opportunity to define individual connections between several tokens. We call these connections „syntagms“ referring to the original meaning of the word. Within the database these syntagms are defined in the following manner: Each token bears an unique identifier. By means of an intermediate table it is possible to connect as many tokens as necessary. Each token can be part of an unlimited number of connections. Each of these connections can itself be tagged on its part. By combining morphological and syntactical criteria search results are much more accurate. On the other hand the definition of syntagms entails additional effort which is the reason why this instrument is only applied if search results based upon purely morphological criteria are not satisfying.



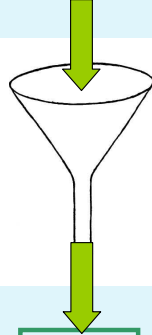
I will now try to explain how the geographical maps are processed.

All the geographical maps containing the variable linguistic informations are processed „on the fly“ by a PHP-Module which is run on the server. When ready, the map is delivered to the Webbrowser running on the client computer. The necessary data are being extracted from our Backend-MySQL-Database.

F11: Ieri la mia macchina si è rotta di nuovo

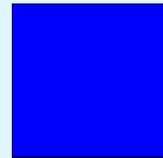
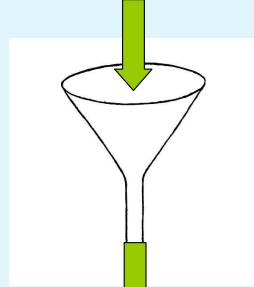


To explain the procedure in detail I will make use of the example of the question who of our informants has used the *passato prossimo* and who the *passato remoto* when formulating the sentence „Ieri la mia macchina si è rotta di nuovo“ (our F11).

passato remoto + F11


Bag2mDQ1	1	rumpiu	ajeri a makina si rumpiu n aṭra vota
Bag2wDQ1	1	rumpiu	ieri a me kamina si rumpiu n aṭra vota
Biv1mDQ1	1	ruppiu	a makkina mia s ruppiu ajeri
Biv1wDQ1	1	ruppiu	ajeri makina mia s ruppiu
Biv1wQ1	3	ruppiu	a makkina s ruppiu
...			

The search strategy looks like this: We tell the database to look for all the tokens that belong to answers to the stimulus F11 and that are connected with types that are tagged as *passato remoto*. We also tell the database to extract the informant's code and the count of the tokens found.

passato prossimo + F11


Bel1mDQ1	4	è	rutta	ajer a makkina mia s	è rutta torna
Bel1mlQ1	1	è	rott	ajer a mmakena s	è rott torna i nnuavu
Bel1wDQ1	4	è	rruth	a makina ajar torna s	è rruth
Bel1wlQ1	1	è	rutta	ajeri a makina s	è rutta nov i ttom i nnua
Bel2mDQ1	1	è	ruttha	ajeri a ma a makkina mia s	è ruttha
...					

Afterwards we look for incidents of the use of the *passato prossimo*. Again the search is limited to tokens connected with the stimulus F11. This time the search is a bit more complicated since there is no specific morphological form to search for. The *passato prossimo* is formed by two words: An auxiliary verb followed by a participle perfect. Normally the two words adjoin directly. Given these preconditions the search has to be extended to two words at a time with the additional requirement that the distance between the two words may not exceed one token at maximum. Again we tell the database to register the names of the informants.



LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

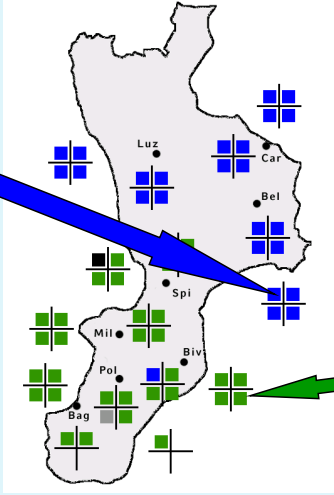
Map processing: Example



passato prossimo

■

- Bel1mDQ1
- Bel1mlQ1
- Bel1wDQ1
- Bel1wlQ1
- Bel2mDQ1
- Bel2mlQ1
- Bel2wDQ1
- Bel2wlQ1
- Biv1mlQ1
- Car1mDQ1
- Car1mlQ1
- Car1wDQ1
- Car1wlQ1
- Car2mDQ1
- Car2mlQ1



passato remoto

■

- Bag1mDQ1
- Bag1mlQ1
- Bag1wDQ1
- Bag1wlQ1
- Bag2mDQ1
- Bag2wDQ1
- Biv1mDQ1
- Biv1wDQ1
- Biv1wlQ1
- Biv2mDQ1
- Biv2mlQ1
- Biv2wDQ1
- Biv2wlQ1
- Mil1mDQ1

luecke@lmu.de

12/13

At this point we have two results: One containing the names of the informants making use of the *passato prossimo*, the other containing the names of the informants using the *passato remoto*.

These two results are then handed over to the PHP-module of the system that is also running on the server. PHP stores the results in a special vectorial variable (a so called „array“) and combines it with another variable containing the coordinates of the symbols representing each informant on the geographical map. Finally the color of each symbol is assigned according to the information whether the informant's name is listed in one or in the other group. In case a speakers name shows up in both results the corresponding symbol is dyed black, if a speakers name doesn't show up in any result – for whatever reason – the symbol is colored grey. At last PHP generates a picture in png-format and hands it over to the webserver which on its part sends it to the browser of the client.

asica.gwi.uni-muenchen.de

www.itg.lmu.de

Thank you!

And now, due to time restriction, I have to finish my lecture and hand over to my colleague, Mrs Salminger. If there should be time and interest I am willing to give additional explanations afterwards. The text and slides of my lecture can be accessed on the project homepage which is accessible under the internet address you can see in the first line on this slide.

Thank you!