



TEXT



GRID

für die Korpuslinguistik

ToC – Andrea Zielinski

- Korpuslinguistik
 - Motivation
 - Fachwissenschaftliche Perspektive
 - Anwendungsszenarien
 - Ausblick

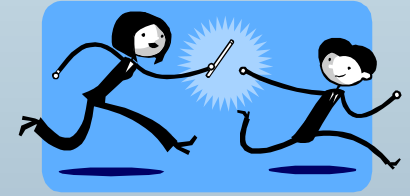


TextGrid ... für die Geisteswissenschaften

- Virtuelle Forschungsumgebung für die “eHumanities“
- Generische Plattform für die wissenschaftliche Textverarbeitung
- Service Grid (Werkzeugkasten für kollaboratives Arbeiten)
- Data Grid (virtuelles Archiv für nachhaltige Datenhaltung)



Ziele von TextGrid



- Kollaborative und interdisziplinäre Arbeitsweisen verstärken
 - den Austausch von Quellendaten fördern
 - gemeinsame Entwicklung von Werkzeugen
 - Integration von relevanten Werkzeugen unter einer Benutzerschnittstelle
 - Integration weiterer Daten/Archive
 - Definition und Konfiguration von Arbeitsabläufen
-

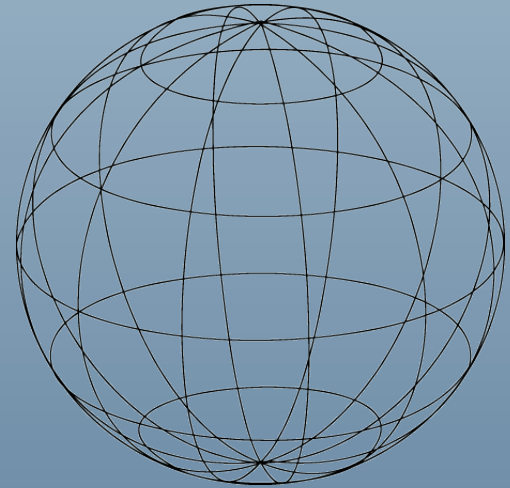
TextGrid ... und die Korpuslinguistik

- Virtuelle Forschungsumgebungen
 - Kooperationen zwischen unterschiedlichen linguistischen Disziplinen
- Generische Plattform für die Textverarbeitung
 - Tools zur Analyse, Annotation und Suche
- Service Grid
 - hohe Verarbeitungszeit für umfangreiche (> 1 Terabyte) Korpora
- Data Grid
 - Integration bisher unverbundener projektspezifischer Daten
 - Ressourcen (z.B. Thesauri, Lexika und Namenslisten) sind auch für andere Fachgebiete interessant

Aktuelle Forschungsthemen:

Grid-basierte Korpuslinguistik:

- Datenrepräsentation:
 - Geeignetes Datenmodell für linguistische Ressourcen
 - Einheitlicher Metadatenstandard für die TextGrid-Community
 - Aktuelle Standards für die Annotation linguistischer Daten
- Modulare WebServices
 - Interoperabilität
 - Basisontologie für NLP-Dienste
 - Orchestrierung von NLP-Diensten
 - Typische Workflows



Metadatenstandards

Paradigmen:

- **Grid & Semantic Web**
RDF, Topic Maps,..
- **Digitale Bibliotheken**
Dublin Core, MARC, METS,..
- **NLP-Technologie**
OLAC, IMDI, TEI, CES/XCES, ISLE,..
- **Computerphilologie**
TEI

Annotationsstandards

Paradigmen:

- **Grid & Semantic Web**
RDF, Topic Maps,..
- **Digitale Bibliotheken**
-
- **NLP-Technologie**
TEI, CES/XCES, ISLE, MATE
- **Computerphilologie**
TEI

Standards in TextGrid

Datenmodell für morpho-syntaktische Annotationen

- Verwendung von Feature-Strukturen
 - Primärebene ist <w>, die Wortebene mit
 - Attribute *lemma* und *ana*

Zeichenkodierung:

- Unicode

Metadaten:

- **Minimales Set für alle Bereiche:** DC (~15 Tags)
- Relevante Tags für die Korpuslinguistik: CES (~100 Tags)

Annotationen:

- **Kernkodierung:** Subset von TEI (~15 Tags)
- Optional: etwas größeres TEI-Tagset mit Erweiterungen (~25 Tags)

“Supporting Text Mining for e-Science: the challenges for Grid-enabled Natural Language Processing”

Einige Fragen von John Carroll; Roger Evans; Ewan Klein

- Is stand-off annotation a suitable basis for representing language data annotation on the Grid?
- Can we make a specific proposal for an annotation model for Grid-enabled NLP applications,
- Do we need more than one (eg for text vs speech), or do we need to allow for application-specific annotation models?
- How visible can/should the annotation model be to the user (or to different classes of user)?
- What is an appropriate framework for describing the configuration and composition of NLP applications?
- What is an appropriate interface between NLP components and other parts of a larger system?
- How important is the ability to abstract over services and workflows, and work with saved templates, and how should it be achieved?
- What actual service components can we identify as being potentially useful, and what are the trade-offs between different decompositions?

“Supporting Text Mining for e-Science: the challenges for Grid-enabled Natural Language Processing”

(John Carroll; Roger Evans; Ewan Klein)

Fortsetzung

- How should we describe the topology of NLP services? How should it reflect the topology of both the algorithm and the dataset?
- What range of provision exists on the Grid, and do we need to enhance what is available with Grid hardware resources specifically tuned to NLP requirements?
- How do we match application requirements to available Grid resources, and how important is the accuracy of the matching process? Can complex NLP Grid applications be delivered in a way that offers near Googlelevel ease of use and response times?
- Are there benefits from doing so, compared with a dedicated Google-like service? (For example, better access and interoperability with other Grid services, or the infeasibility of setting up a dedicated service.)
- If not, what would be acceptable ease of use and response time levels for (at least) typical e-Science users?

Anwendungsszenarien

- Typische Arbeitsschritte bei der Erstellung von Korpora
 - Metadatenerfassung
 - Strukturelle Basisauszeichnung
 - Linguistische Auszeichnung
 - Analyse
 - automatisch/manuell/halbautomatisch
 - sequentiell/parallel
 - basierend auf Grammatiken, Wörterbüchern
 - Auswertung (einzeln/kollaborativ)
 - Ergebnis (eindeutig/ambig/konkurrierend)
- Suche und statistische Auswertungen (quantitativ/qualitativ)
 - im Primärtext
 - auf den Annotationen
 - anhand der Metadaten

Beispiel: Metadatenerfassung

Quellenverzeichnis des
Verbunds Mittelhochdeutscher
Wörterbücher mit
inhaltlichen Metadaten
(Klassifizierung nach Raum,
Zeit und Textsorte),
hier 'Nibelungenlied'

The screenshot shows a software window titled 'QVZ' with a search bar at the bottom. The main content area displays two entries for 'Nibelungenlied'. The first entry is for 'Nib.' and includes a detailed description, edition information, and a classification box. The second entry is for 'Nib. B' and 'Nib. C', including a description of the variants and another classification box. The interface has a yellow and blue color scheme and a vertical toolbar on the right with letters A-Z.

QVZ

NIB., Nib.

[Nibelungenlied (nach Hs. A)]. In: Der Nibelunge Noth und die Klage. Nach der ältesten Überlieferung mit Bezeichnung des Unechten und mit Abweichungen der gemeinen Lesart hg. von K. Lachmann. 3. Ausg. Berlin 1851. [5. Ausg. Berlin 1878 (Neudr. Berlin 1960)]
[Lachmanns Zählung bei Bartsch/de Boor *rechts* am Strophenrand]
D: wohl zw. 1190 und 1200
Lexer benutzt zusätzlich:
Wörterbuch zu der Nibelunge Not [nach Hs. A, mit Beiträgen aus anderen Hss.]. Von A. Lübben. 2., verm. und verb. Aufl. Oldenburg 1865. [3. Aufl. Oldenburg 1877 (Neudr. Walluf 1966)]

Klassifikation in der Datenbank

Name: Nibelungenlied (nach Hs. A), ed. Lachmann
Textsorte: Epische Großformen => Heldeneplik aus heimischer, germanischer Tradition
Region: Oberdeutsch => Bairisch => Südbairisch
Entstehungszeit: 1190 - 1200

Nib. B
(mit Strophenzahl), Nib. C
(mit Strophenzahl)

Hs. B bzw. C des Nibelungenlieds, offensichtlich zitiert nach Lachmanns - als selbständiges Werk erschienenem - Variantenapparat zu seiner Ausgabe: K. Lachmann: Zu den Nibelungen und zur Klage. Anmerkungen. Berlin 1836.
[Zählung der B- und C-Strophen jeweils nach A! In der Ausg. von Bartsch/de Boor (nach B) findet sich diese A-Zählung am rechten Rand, in Hennigs Ausg. (nach C) rechts in Klammern]

Klassifikation in der Datenbank

Name: Nibelungenlied, Varianten der Hs. B und C aus Lachmanns Ausg.
Textsorte: Epische Großformen => Heldeneplik aus heimischer, germanischer Tradition
Region: Oberdeutsch => Bairisch => Südbairisch
Entstehungszeit: 1190 - 1200

Text: Suchen

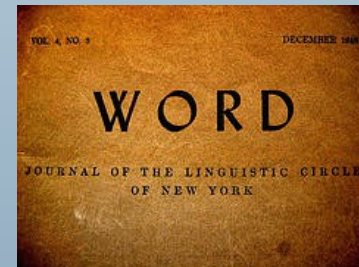
Beispiel: Basisauszeichnung in TEI

TEI-Auszeichnung
Für einen Zeitungstext aus Erfurt,
1769
(Quelle: GerManC)

```
51 | <p><s>Bey <rs type="person" key="HJ">dem ju&#868;ngern Hansy</rs>: <hi rend="antiqua"><title>Hist
52 | zung mit Vergnu&#868;gen lesen;</s> <s>nur den vo&#868;lligen Geist des Werks und die ganze <fore
53 | </p></divl>
54 | <divl type="article"><head><name type="place" key="Z">Zu&#868;rich</name>.</head>
55 | <p><s>Bey <name type="org" key="OGCo">Orell, Ge&#223;ner und Compagnie</name>: Die Grunds&#868;t
56 | </p></divl><fw type="footer"><name type="place" key="RI">Rinteln</name>.</fw><pb n="235"/><fw type
57 | <divl type="article"><head><name type="place" key="RI">Rinteln</name>.</head>
58 | <p><s>Folgenden Aufsatz, den man uns zugeschickt hat, ru&#868;cken wir ein, ohne an der Streitigk
59 | </p><p><s>Herr <rs type="person" key="MJDM">M. J. D. Mu&#868;ller</rs>, <foreign lang="lat"><hi r
60 | mathematische <foreign lang="lat">Methode</foreign> gewo&#868;hnet ist.</s> <s>Um zur Sache selbs
61 | ton dahinreissen lassen, und die mathematische Lehrart verachten, welche doch ein <foreign lang="
62 | </p></divl>
63 | <divl type="article"><head><name type="place" key="LE">Leipzig</name>.</head>
64 | <p><s><title>Versuch u&#868;ber die Geschichte der bu&#868;rgerlichen Gesellschaft</title> von <n
65 | unter den Menschen, von den Grunds&#868;tzen des Kriegs und der Zwietracht, von den Kra&#868;ft
66 | </p><p><s>Der <num type="ordinal" value="2">zweete</num> Theil erza&#868;hlt die Geschichte roher
67 | </p><p><s>Die Geschichte der Staatskunst und anderer Ku&#868;nste ist der Gegenstand des <num typ
68 | </p><p><s>Im <num type="ordinal" value="5">fu&#868;nften</num> und <num type="ordinal" value="6">
69 | </p></divl>
70 | <divl type="article"><head><name type="place" key="LE">Leipzig</name>.</head>
71 | <p><s>Bey <name type="org" key="Jun">Junius</name>: <title>Versuch in <foreign lang="fra">moralis
72 | </p></divl>
73 | <divl type="article"><head><name type="place" key="Z">Zu&#868;rich</name>.</head>
74 | <p><s>Bey <name type="org" key="OGCo">Orell, Ge&#223;ner und Compagnie</name>: <rs type="person"
75 | sten Fru&#868;chte einer gesunden Philosophie zu za&#868;hlen ist.</s> <s>Mit dem Auge des Beobac
76 | </p></divl>
77 | <divl type="article"><head>Gelehrte Nachrichten.</head>
78 | <p><s>Zu <name type="place" key="Je">Jena</name> sind die Herren Doctoren <rs type="person" key="
79 | </p><p><s>Man meldet uns aus <name type="place" key="B">Berlin</name>, da&#223; die <title>Oden</
```

Beispiel: Morpho-syntaktische Auszeichnung

morpho-syntaktische Annotation im American National Corpus (ANC)
Stand-off Kodierung, basierend auf XCES



```
<p id="p3"><s id="p3s1">
```

```
Ireland has been inhabited since very ancient times, but Irish history really  
begins with the arrival of the Celts around the 6th century b.c.</s>
```

```
<chunk type="sentence" xml:base="#p3s1">
```

```
<tok xlink:href="xpointer(string-  
range(' ',0,7))"><msd>np++++</msd><base>ireland</base></tok>
```

```
<tok xlink:href="xpointer(string-  
range(' ',8,11))"><msd>vbz+hvz+aux++</msd><base>have</base></tok>
```

```
<tok xlink:href="xpointer(string-  
range(' ',12,16))"><msd>vprf+ben+aux+xvbnx+</msd><base>be</base></tok>
```

Beispiel: Suche in Korpora

KWIC-Ausschnitt
zum Adjektiv *frei*
mit dem
Recherchetool
'COSMAS'

The screenshot shows the COSMAS search interface in a Microsoft Internet Explorer browser window. The title bar reads "COSMAS Konkordanzen - Microsoft Internet Explorer". The main content area displays search results for the word "frei".

At the top, there are navigation tabs: "Verfassen", "Korpus", "Suchen", "Ergebnis", "Kwic", "Volltext", and "Export". Below these, the search parameters are shown: "Korpus: wiv Korpus des Projektes Wissen über Wörter" and "Suchanfrage: &frei (10041 aus 290657 Treffern ausgewählt)".

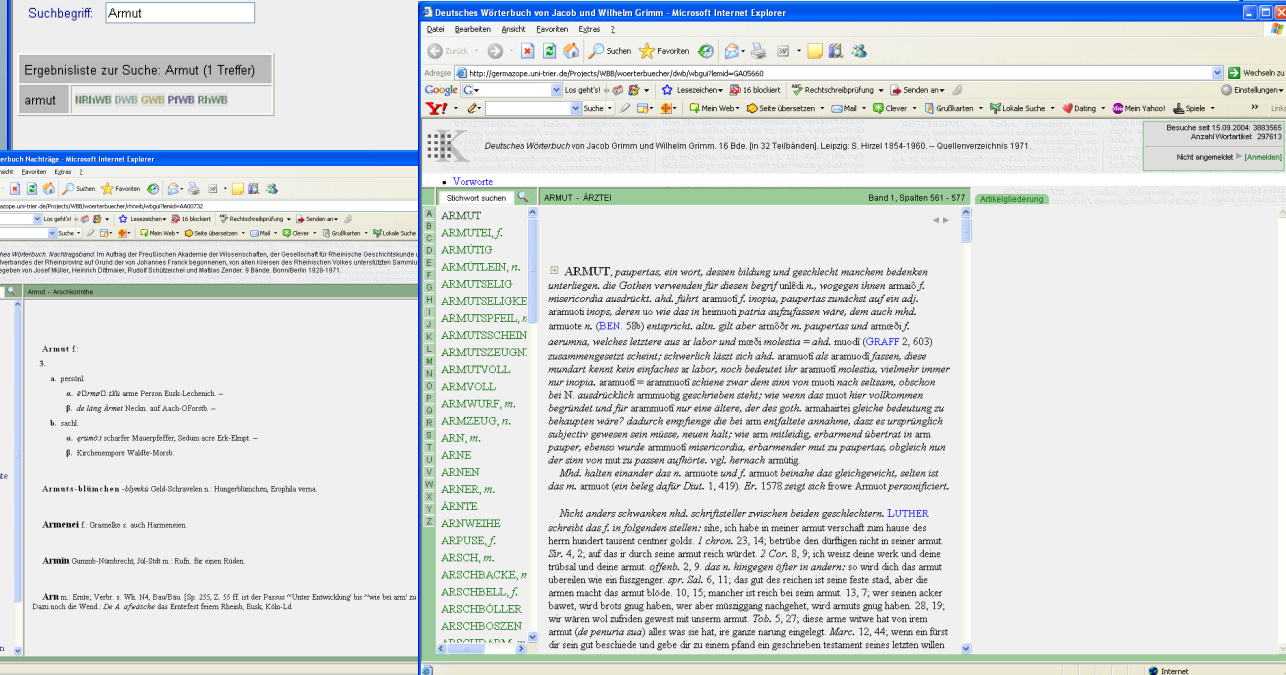
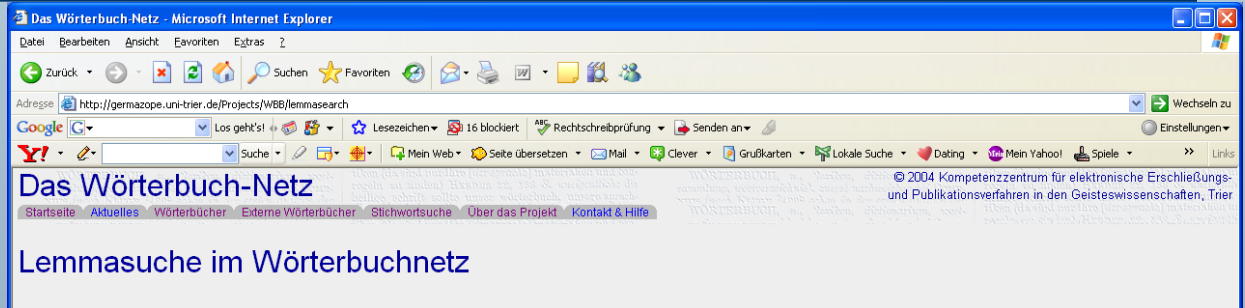
The search results are displayed in a table-like format with columns for "links", "rechts", and "Sätze". The "links" column shows the number of occurrences to the left of the search term, and the "rechts" column shows the number to the right. The "Sätze" column shows the full sentence containing the search term.

The search results are grouped by date and source: "B97/710, Berliner Zeitung, Oktober 1997", "B97/711, Berliner Zeitung, November 1997", and "B97/712, Berliner Zeitung, Dezember 1997".

Each result entry consists of a list of checkboxes followed by a sentence. The search term "frei" is highlighted in the original image. For example, in the first result, the sentence is: "uch. Acht Versuche hatte jeder Kandidat frei, was in der Summe 96 unbedrängte He".

Beispiel: Suche in Wörterbüchern

Lemmaintrag für
Armut
im Trierer 'Mhd.
Wörterbuchnetz'



Vorgehensweise TextGrid

- Sichtung
 - Tools verschiedener relevanter Projekte aus einschlägigen Journalen
- Aufbau eines Testbed, Input-/Output-Werten für Tools
- Codierung
 - Standardisierung auf der Basis von TEI, UNICODE ...
 - Standards (DC, W3C etc.)
 - „TEI-Kernkodierung“ spiegelt homogene Basis-Suchanfragen
UND „Spezialcodierung“ spiegelt Tiefenauswertung
 - Corpus-übergreifende Suche (cross-search)
- Austausch mit der Community zwecks Schnittstellendefinition

Chancen: Wörterbuchredaktionssystem für historische Sprachstufen

Deutsches Fremdwörterbuch Ausschnitt 'Fidibus'

Fidibus M. (-, auch Fidibusses; -, auch Fidibusse), im späteren 17. Jh. aufgekommene latinisierende Bildung unsicherer Herkunft, wohl nach lat. *fidibus*, Ablativ Pl. von *fides* 'Saite, Saitenspiel, Saiteninstrument, Lyra, Leier' (sofern tatsächlich, wie Moriz Haupt vermutete, hier eine witzige Umdeutung einer Horazstelle vorliegt: in Od. 1,36,1-3 *et ture et fidibus iuvat placare .. deos* 'lasst uns mit Weihrauch und Saitenspiel die Götter besänftigen' wären demnach scherzhaft *ture* (Ablativ Sing. von lat. *tus* 'Weihrauch') als 'Tabaksqualm' und *fidibus* (s.o.) als 'Pfeifenanzünder' interpretiert worden; daneben wird die Herkunft des Wortes auch anders erklärt, s. Belege 1722, 1738, 1770), früher auch in der Schreibung *Vidibus* und in den entstellten Formen *Vidimus* und *Fiedepuß*.

Zunächst studenten- und bildungsspr., dann ugs. und in neuerer Zeit mit der Sache zunehmend ungebr. in der Bed. '(gefalteter, geknickter oder gerollter) Papierstreifen zum Anzünden einer Tabakspfeife, Zigarre usw.', in Wendungen wie **ein brennender Fidibus, frische Fidibusse, einen Fidibus rollen/anzünden/ausklopfen, ein Bund Fidibus zurechtschneiden** und Zss. wie **Fidibusbecher, -schnitzel; Tabak-, Knallfidibus** 'aus Knallpapier hergestellter Scherzartikel'; öfter auch in der abwertenden Wendung **eine Zeitung, Schrift als Fidibus gebrauchen, aus einem Buch/Roman/Brief Fidibus verfertigen** 'ein schlechtes, wertloses Schriftstück als Anzünder verwenden und dadurch vernichten', von daher vereinzelt in der Bed. 'wertloses Geschreibsel, Schund' (→ Makulatur), in Wendungen wie **nichts als Fidibus schreiben** (s. Beleg 1769).

Chancen: Translation Memory für die Übersetzung

Ausschnitt aus dem
Englisch-Französisch
Hansard Korpus
(CES-kodiert)

Alignment of - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Zurück Suchen Favoriten

Adresse <http://tali.iro.umontreal.ca/Ressources/BAF/corpus/hans.fr-en.html> Wechseln zu

Google C Los geh'st! Lesezeichen 16 blockiert Rechtschreibprüfung Senden an Einstellungen

Suche Mein Web Seite übersetzen Mail Clever Grußkarten Lokale Suche Dating Mein Yahoo! Spiele Links

Initiatives parlementaires	Private Members' Business
CHAMBRE DES COMMUNES	HOUSE OF COMMONS
Le lundi 14 mars 1994	Monday, March 14, 1994
La séance est ouverte à 11 heures.	The House met at 11 a.m.
Prière	Prayers
[Français] Loi sur le Parlement du Canada	[Translation] Parliament of Canada Act (oath or solemn affirmation)
Projet de loi C-201. Motion visant à la deuxième lecture	Bill C-201. Motion for second reading
--Madame la Présidente, je vous remercie de me permettre de présenter le projet de loi C-201, le premier projet de loi émanant des députés à être présenté à cette session. Il aborde la question de l'assermentation que nous faisons à la Reine, auquel j'aimerais apporter une modification.	He said: Madam Speaker, I welcome this opportunity to introduce Bill C-201, the first Private Members' bill of this session. It concerns the oath of allegiance we swear to the Queen, to which I would like to make some changes.
[Traduction] Le 9 novembre 1993, dans le cadre de ma cérémonie d'assermentation, j'ai eu le privilège en tant que député élu du Parlement de prêter le serment d'allégeance à la reine.	[English] On November 9, 1993 at my swearing in ceremony I had the honour as an elected member of the Canadian Parliament of pledging allegiance to the Queen.
Après avoir été élu à la Chambre par les gens de ma circonscription de Carleton--Gloucester et avoir obtenu à cette occasion un chiffre record de 46 800 voix, soit environ 35 000 de plus que mon plus proche adversaire, je me suis senti très fier, mais par-dessus tout flatté d'avoir été choisi pour représenter un si grand nombre de Canadiens. C'est pourquoi je voulais ajouter au serment d'office, c'est-à-dire celui dans lequel nous promettons allégeance à la reine, une déclaration de loyauté envers le Canada et sa constitution.	Having been elected to Parliament by the electors of my riding of Carleton--Gloucester by a record 46,800 votes in my favour, about 35,000 votes more than my nearest challenger, I felt proud but above all I felt honoured of having been elected to serve so many Canadians. For this reason I want to add to the present oath of office, that is to say the one that pays allegiance to the Queen, a pledge of allegiance to Canada and its Constitution.
Après avoir juré de servir la reine sur ma bible de famille et avoir signé les documents parlementaires que m'a remis le greffier de la Chambre des communes, en présence de mon épouse et de mes enfants, je lui ai demandé de me laisser lire la déclaration suivante:	After swearing allegiance to the Queen on my family Bible and signing the parliamentary documents handed to me by the Clerk of the House of Commons in the presence of my wife and children, I requested that the Clerk of the House of Commons let me read the following affirmation.
	[Translation]

Fertig Internet

Chancen: Aufbau eines umfangreichen Deutschen Diachronen Korpus

Diachrone Untersuchungen zum Sprachwandel:

- Formwandel
Beispiel: *frouwa* im Althochdeutschen, *vrouwe* im Mittelhochdeutschen, *Dame* im Neuhochdeutschen
- Bedeutungswandel
Beispiel: Bedeutung von *Karriere* im Althochdeutschen 'schnellster Galopp', 'voller Lauf', 'höchste Geschwindigkeit'
Beispiel: Projekt Klassikerwortschatz (Freiburg)

Ziel: Normalisierung sprachstufenbezogener Unterschiede

→ Generierung einer Metalemmaliste

Varianten-, Kookkurrenz- und Kollokationsanalyse

Chancen: Recherchetool mit intelligenter Suche (Phrasen, Idiome, NE)

Motivation für die Indexierung mit Phrasen:

- Termzusammenhang, aber nicht benachbart

Anschwellen der auf dem Foto sichtbaren Bauchspeicheldrüse

- Benachbart, aber kein Termzusammenhang

In der exokrinen Drüsenfunktion werden vom endokrinen Drüsenanteil Hormone direkt ins Blut abgegeben

- Keine Abstraktion über syntaktische Varianten

*Anschwellen der Bauchspeicheldrüse
angeschwollene Bauchspeicheldrüse*



Zusammenfassung & Ausblick

- Aufbau einer Community in den Geisteswissenschaften mit europäischer/internationaler Vernetzung (Virtual Research Environment, e-Research)
- Kooperationen in Deutschland:
 - eSciDoc, Text-Archive ...
(MPDL: Tools, Repositorien, AstroGrid: Metadaten, Stemnet/PharmaGrid: Computerlinguistik + Textextraktion)
- Kooperationen auf EU-Ebene, e-Humanities:
 - Assoziiert mit CLARIN
 - DARIAH im Rahmen von ESFRI (DANS, ADHS, MPDL ...)
 - Interedition: Niederländische Akademie der Wissenschaften (Kollationierung)

Fragen