Wintersemester 2006 • Institut für Germanistik I

Vorlesung
Computerphilologie

Themenfeld Standards

Ist in der Computerphilologie noch alles frei Schnauze oder gibt es schon Standards?

„The nice thing about standards is that there are so many of them to choose from“

Andrew S. Tanenbaum

Übersicht

Es gibt bereits eine große Zahl von Standards, die aus der Computerlinguistik, aus der HLT (Human Language Technology) und aus der Industrie kommen. Sie einzuhalten ist aus Gründen der Vergleichbarkeit und Weiternutzung vernünftig.

Detailfragen:

- Welche Ebenen in Sprach- und höherer Textverarbeitung können / sollten standardisiert werden?
- Was soll / sollte auf diesen Ebenen standardisiert werden?
- Welche Standards gibt es auf welchen Gebieten?

Ebenen von Sprachstandards

- Korpus-Standards
- Test-Standards

- Phonetische Daten-Standards
- Lexikon-Standards
- Text-Standards
- Diskurs-Standards
- Formalismus-Standards

- Software-Standards
- System-Standards

Korpus-Standards

- Korpusstandards werden benutzt für
  - die Statistik sprachlicher Phänomene
    (rein linguistisch, verschiedene Domänen, ...)
  - Test-Suiten
  - Regelanordnung in Grammatiken, Parsern o.ä. Prozessen
- Standardisiert sind
  - Inhalte (häufig aufgabenbezogene Ausschnitte aus Domänen)
  - Codierung (z.B. für Pausen, extraling. Objekte)
- Labelling gehört zu grammatischer Standardisierung
- Für speech-Corpora siehe unter Speech-Standards

Test Standards

Korpus-Standards
Test-Standards
Phonetische Standards
Lexikon-Standards
Text-Standards
Diskurs-Standards
System-Standards
Formalismus-Standards
Software-Standards

- Verhältnis von Testset : Trainingset
- Testintervalle
- Zählverfahren oder Statistische Verfahren
- Dokumentationsstandards
- Tools

Phonetische Daten-Standards

- *Z.B.*
  - *Aufnahmehardware-Standards,*
  - *Fileformat-Standards,*
  - *Headerstruktur und*
  - *Annotationsstandards.*
- *Wichtig für Austausch und Weiterbenutzung von Aufnahmematerial.*

Lexikon-Standards

Korpus-Standards
Test-Standards
Phonetische Standards
Lexikon-Standards
Text-Standards
Diskurs-Standards
System-Standards
Formalismus-Standards
Software-Standards

- Zur Wiederbenutzung und Weitergabe mindestens
  - Wortklassendefinitionen
  - Morphologische Annotationen
  - Generelle Notationsfestlegungen
- Am besten TEI-ähnliche Festlegungen

Text-Standards

Korpus-Standards
Test-Standards
Phonetische Standards
Lexikon-Standards
Text-Standards
Diskurs-Standards
System-Standards
Formalismus-Standards
Software-Standards

- Wiederbenutzung und Austausch mindestens durch standardisierte
  - Header
  - Codierung von
    - Wörtern und nichtlinguistischen Entitäten
    - Phrasen
    - Supersegmentale Phänomene (Diskurs, Kohärenz, turn taking. Dokumentstruktur)
- Am besten durch Übernahme bestehender Standards, wie TEI

Diskurs-Standards

- Dialogstrukturierung
- Spontansprachliche Phänomene
- Turn Phänomene (z.B. Überlappung von Signaldaten)
- nichtlinguistischen Entitäten (direkte deiktische Bewegungen, Diskurssituation, Umgebung, Lärm)

System- und Tool-Standards

Korpus-Standards
Test-Standards
Phonetische Standards
Lexikon-Standards
Text-Standards
Diskurs-Standards
System-Standards
Formalismus-Standards
Software-Standards

- Funktionale Spezifikation
- Prioritätenliste auf der funktionalen Spezifikation
- Robustheit
- Adaptivität
- Wartbarkeit
- Handbücher und Dokumentationen
- Integration in den Arbeitsfluß

Formalismus-Standards

- Codierung auf jeder Ebene, angefangen von Klammerausdrücken
- Benutzung von Mark-up Sprachen, wie SGML, XML oder TEI
- Exchange formats
- Lingua - franca - Definition für die Repräsentation in mehr-Ebenen-Systemen

Software Standards

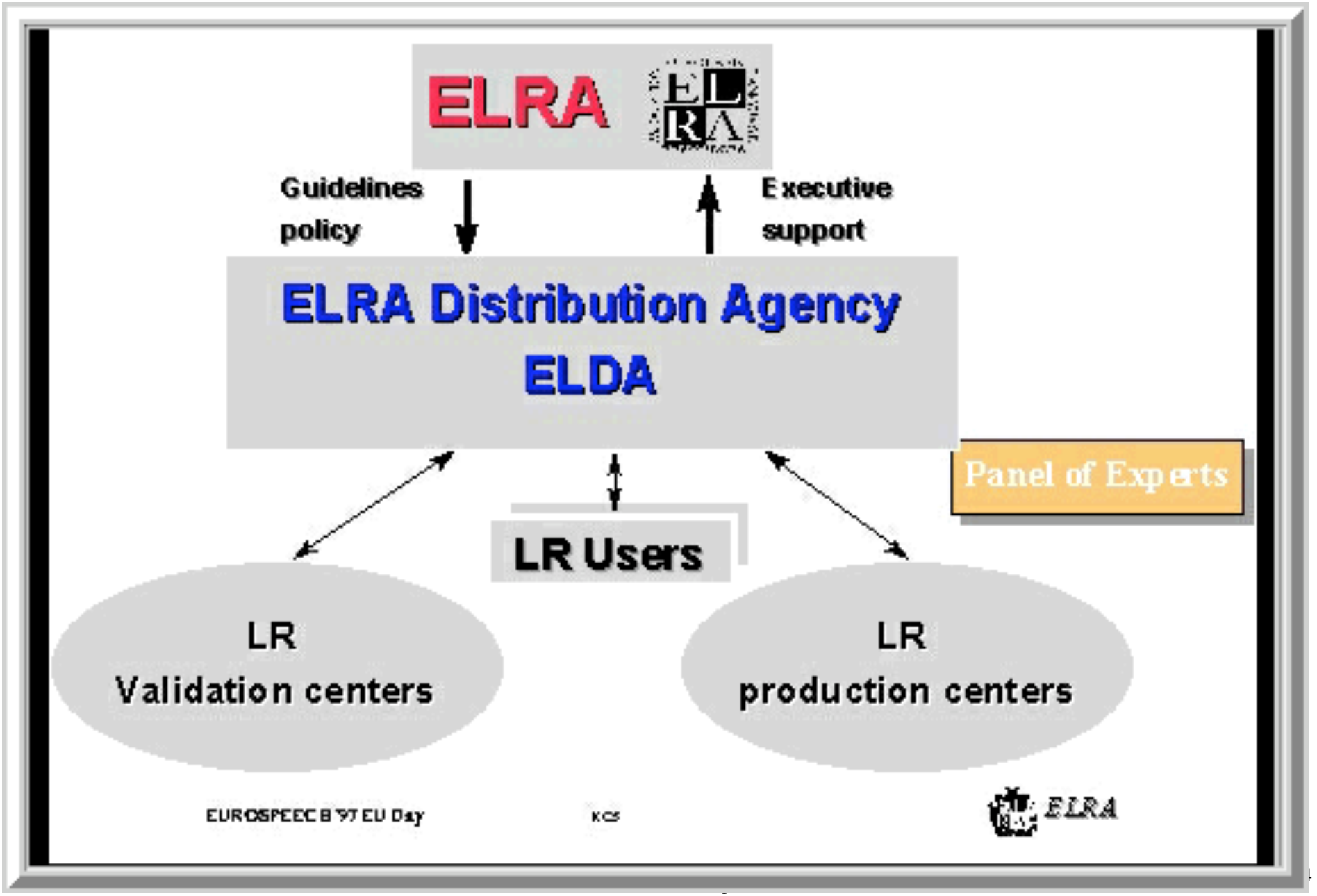
- Abgeleitet von üblichen Software-Qualitätsstandards, wie Durchschaubarkeit, Robustheit, Wartbarkeit, Zuverlässigkeit, Effizienz, etc.
- Besonders wichtig bei textuellen Funtionen in einer Programmiersprache, wie z.B. DCGs (Definite Clause Grammars) oder lexikalische Datenbanken in der Sprache Prolog, da hier der Programmierstil über die linguistische Ausdrucksmächtigkeit entscheidet.

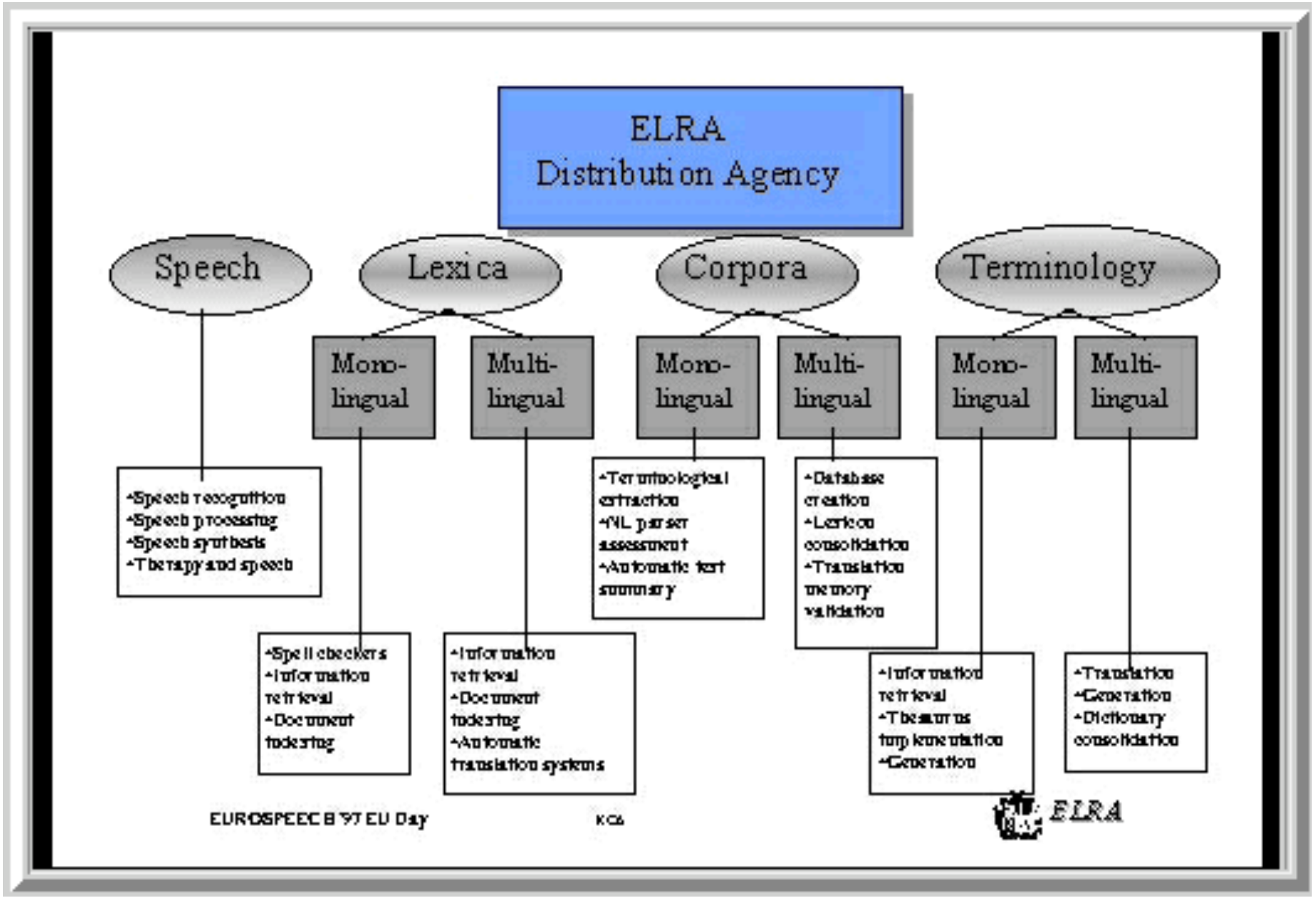
Wichtige Standardisierungsquelle: ELRA

European Language Resources Association

„The European Language Resources Association (ELRA) was established as a non-profit organization in Luxembourg in February, 1995. The overall goal of ELRA is to provide a centralized organization for the validation, management, and distribution of speech, text, and terminology resources and tools, and to promote their use within the European telematics R&TD community“.

- http://www.icp.grenet.fr/ELRA/fr/home.html


ELRA
Guidelines
policy
Executive
support
ELRA Distribution Agency
ELDA
Panel of Experts
LR Users
LR
Validation centers
LR
production centers
EUROSPEECH '97 EU Day
ELRA


ELRA
Distribution Agency
Speech
Lexica
Corpora
Terminology
Mono-lingual
Multi-lingual
Mono-lingual
Multi-lingual
Mono-lingual
Multi-lingual
*Speech recognition
*Speech processing
*Speech synthesis
*Therapy and speech
*Spell checkers
*Information retrieval
*Document tagging
*Information retrieval
*Document tagging
*Automatic translation systems
*Terminological extraction
*NL parser assessment
*Automatic text summary
*Database creation
*Lexicon consolidation
*Translation memory validation
*Information retrieval
*Thesaurus implementation
*Generation
*Translation
*Generation
*Dictionary consolidation
EUROSPEECH '97 EU Day
ELRA

Language Resources of ELAN.

„ELAN will provide standardised resources of the following languages: Belgian French, Bulgarian, Catalan, Czech, Danish, Dutch, English, Estonian, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portu-guese, Romanian, Russian, Serbian, Slovakian, Slovene, Swedish, Turkish and Ukrainian."

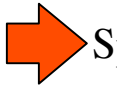
„They will comprise textual resources (corpora) ranging from 1 - 4 million words or more for each language, and lexical resources (several kinds of lexicons) ranging from 5000 to 20 000 entries or more for each language."

Das Projekt war sichtbar unter:

http://solaris3.ids-mannheim.de/elan/

Ist aber zur Zeit nicht (mehr) präsent

Speech Recording Standards



„There are two standards for speech recordings :

- a first one proposed by the European SAM project used to record EUROM1 databases

  (see http://fer.icp.inpg.fr/ICP/bd-europ.html) and now the speechdat projects (see www.speechdat.org/).
- A second one has been proposed by NIST and is used by LDC (see http://www.ldc.upenn.edu/).

  See the EAGLES handbook of standards“

Jeff ALLEN (ELDA) by e-mail

NIST

„You can get NIST's SPHERE software package for speech file manipulation (and other goodies) at:

http://www.itl.nist.gov/div894/894.01/software.htm

This package includes files documenting the SPHERE speech file format, which many sites -- including the LDC -- have adopted.“

File Formats

Speech-Standards

„If you want your material to be easily accessible to a wider community, then a standard like .wav would be a much better choice than the SAM or NIST standards, which are not understood by most commercial audio or multimedia programs.

There are common acoustic file formats such as those represented by the .wav, .aiff, and .au file extensions. You can learn about these in Guido van Rossum's annotated list of audio file formats, which you can find in html version at

ftp://ftp.cwi.nl/pub/audio/index.html“

(M. Liberman, by e-mail)

Realistic Speech

Speech-Standards

In general, most speech corpora represent some particular prototypical task -- such as talking with a stranger on a certain topic -- and the recordings try to model that task closely, warts (e.g. crosstalk on telephone lines) and all. A very few corpora -- TIMIT comes to mind -- have attempted to record speech with no disfluencies and no noise.

(Bill Fisher / NIST)

IAMT Zertifizierung von MT Systemen

- „It must be noted briefly but seriously that the point of the Certification is to provide useful guidance to non-expert users buying software. It is not a research topic for experts and others building or studying MT technology.
- Furthermore, the IAMT cannot allow itself to become partisan in the wars of commerce. Therefore the Certification cannot be an evaluation; evaluation is too difficult and controversial, and too easily slanted toward one or another application. In summary, the Certification has to be simple, fair, and cheap to perform“.

IAMT Zertifizierungs Gremium

Eduard Hovy, USC Information Sciences Institute, Los Angeles, CA (chair)

Laurie Gerber, recently of SYSTRAN, La Jolla, CA

John Hutchins, current President of IAMT, Norwich, England

Sharon O'Brien, ALPNET, Washington, DC

John O'Hara, recently of Lernout & Hauspie, San Diego, CA

Joerg Schuetz, IAI, Saarbruecken, Germany

Muriel Vasconcellos, MTNI Newsmagazine editor, San Diego, CA

John White, Litton PRC, Washington, DC

EAGLES (aus der Website)

Allg. Standards

- The Expert Advisory Group on Language Engineering Standards (EAGLES) is an initiative of the European Commission, within DG XIII Linguistic Research and Engineering programme, which aims to accelerate the provision of standards.
- Numerous well-known companies, research centres, universities and professional bodies across the European Union are collaborating to produce the
- EAGLES Guidelines which set out recommendations for de facto standards and for good practice in the above areas of language engineering.

Standardisierungsobjekte von EAGLES

- Very large-scale language resources (such as text corpora, computational lexicons and speech corpora);
- Means of manipulating such knowledge, via computational linguistic formalisms, mark up languages and various software tools;
- Means of assessing and evaluating resources, tools and products.

General Standards

EAGLES Websites

- http://www.linglink.lu/le/projects/eagles/ann967.html &
- http://www.ilc.pi.cnr.it/EAGLES/home.html)

Speech Daten Handbuch

Speech-Standards

- EAGLES handbook of standards and resources for Spoken Language Systems by Dafydd Gibbon, Roger Moore and Richard Winski (eds) , Mouton de Gruyter, and there are web sites on EAGLES work

EAGLES Evaluation von NL Systemen

Status	Part	Title
	Chapter 1	Reading Guide
P	Chapter 2	The Framework Model
B/D	Chapter 3	Related Work
	Chapter 4	Conclusions and Directions for Future Work
	References	
	A.	ISO Terms and Guidelines
	B.	Requirements Analysis for Linguistic Engineering Evaluation
	C.	Evaluation of Writers' Aids
	D.	Evaluation of Translators' Aids
	E.	User Profiles
	F.	Translation Tools
	G.	Feature Checklist Examples
	H.	A Practical Evaluation of Writers' Aids
	I.	Evaluation of Knowledge Management Systems
	J.	SdT: A Case Study
	Appendices	
	Bibliography	

Document Status Labels

General Standards

- R: recommendations
- P: preliminary recommendations
- F: formal specifications and explicit guidelines
- V: validation document
- B: background document
- D: data produced according to EAGLES recommendations
- L: links to related project documents

Text Standards

Text Corpora

Status	Title	FTP version	Last updated	Edited for release
Diagram	*Reading Guide*	corpintr.ps	071296	YES
P	*Preliminary recommendations on corpus typology*	corpustyp.ps		NO
P	*Preliminary recommendations on text typology*	texttyp.ps	071296	YES
R/F	*Recommendations on corpus encoding*	ces.tar	101496	YES
R	*Recommendations for the morphosyntactic annotation of corpora*	annotate.ps	071296	YES
P/B	*Preliminary recommendations on syntactic annotation of corpora*	sasg1.ps sasgappa.ps sasg2.ps sasg3.ps	071296	YES
P	*Preliminary recommendations on spoken text corpora*	spokentx.ps		NO
B	*Reusability of linguistic software*	lsd1.tar	042896	NO
B	*Guidelines for Linguistic Software Development*	lsd2.tar	042896	NO

Computational Lexicons

Lexicon Standards

Status	Title	FTP version	Last updated	Edited for release
Diagram	*Reading Guide*	guide.ps	071296	YES
R	*Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora A common proposal and applications to european languages*	morphsyn.ps		YES
P	*Preliminary recommendations on subcategorization*	synlex.ps synlexap.ps	101896	YES
F	*ELM-DE: E4 GLES specifications for German morphosyntax*	elm_de.ps	071596	NO
V F	*E4 GLES validation task on tagset mapping ELM-EN: E4 GLES specifications for English morphosyntax (Appendix)*	map_en.ps		NO
F	*ELM-IT: E4 GLES specifications for Italian morphosyntax*	elm_it.ps		NO
F	*ELM-FR: E4 GLES specifications for French morphosyntax*	elm_fre.ps	050697	YES
V	*Study of the relation between tagsets and taggers*	tags.ps		NO
B	*Lexicon architecture*	lexarch.ps		NO
B	*Computational lexicons methodology task*	method.ps		NO
P	*Preliminary Recommendations on Semantic Encoding*	rep4.ps	280299	YES

Computational Linguistics Formalism

Status	Title	FTP version	Last updated	Edited for release
Diagram	Ch 1: Reading Guide	fwg-report.ps	071296	YES
B	Ch 2: State of affairs at the beginning of EAGLES			
B	Ch 3: New developments since the start of EAGLES			
B	Ch 4: Trends towards convergence			
B	Ch 5: Reuse of grammars			
P	Ch 6: Needs for industry and research			
P	Ch 7: Recommendations concerning formalisms for new NL projects			
P/F	Ch 8: Exchange formats for grammar formalisms			
	Ch 9: Conclusion and planned activities			
	Bibliography			
B	*Report of the EAGLES Workshop on Implemented Grammar Formalisms*	grammar.ps		NO
B	*Report of the EAGLES Workshop on Linguistic Adequacy of Linguistic Formalisms for NLP*	adequacy.ps		NO
B	*Report of the EAGLES Workshop on the Connection between CLP and Linguistic Formalisms*	clpling.ps		NO

Kontrollierte Sprachen - Definition -

Text-Standards

Kontrollierte Sprachen sind eine Untermenge von natürlichen Sprachen, die meist beschränkt sind in

- Wortschatz und
- Grammatik

Kontrollierte Sprachen werden benutzt, um die Ambiguität und Komplexität natürlicher Sprachen zu reduzieren oder ganz auszuschalten.

Kontrollierte Sprachen
- Zweck -

Text-Standards

2 Kategorien:

- **Mensch-orientierte** (HOCL - Human Oriented Controlled Language). Sie versuchen die Lesbarkeit von Texten für Menschen zu erhöhen.
- **Computer-orientiert** (MOCL- Machine Oriented Controlled Language). Sie versuchen die Verarbeitbarkeit durch einen Computer zu erhöhen (meistens benutzt für maschinelle Übersetzung).

Kontrollierte Sprachen (HOCL) - Beispiele -

Text-Standards

- AECMA Simplified English (SE) - Standard für technische Dokumentationen im Flugzeugbau.
- Easy English - entwickelt für Menschen, die Englisch als Fremdsprache sprechen (erster Zweck: “Übersetzung” der Bibel vom Englischen in Easy Englisch!)

AECMA Regeln

- Do not use forms of the verb not shown in the dictionary (such as verbs in the ‚-ing' form)
  - Example: *when the waste burns, ...*
  - Instead of: *when the waste is burning...*
  - Approved forms: *Burn, burns, burned, burned*
- Use approved words from the dicitionary only as the part of speech given
  - Example: *test* is a noun (and not a verb)

Beispiele aus dem AECMA Lexikon

- Approved (adj.): Permitted by „certification“
- Area (n): A surface in specified limits
- Around (pre): On all sides
- Arrangement (n): *configuration*
- Article (n): Use: *object*
- Ask (v) Use: *tell, speak*
- Assign (v): Use: *give*
- Attendence (n): Use: *be*

Kontrollierte Sprachen (MOCL) - Beispiele -

Text-Standards

- Alcatel's COGRAM
- IBM's Easy English
- Sun Microsystem's Controlled English

- Es gibt auch kontrollierte Sprachen für Deutsch, Schwedisch, Griechisch und Spanisch.

Beispiele für häufige Regeln in kontrollierten Sprachen

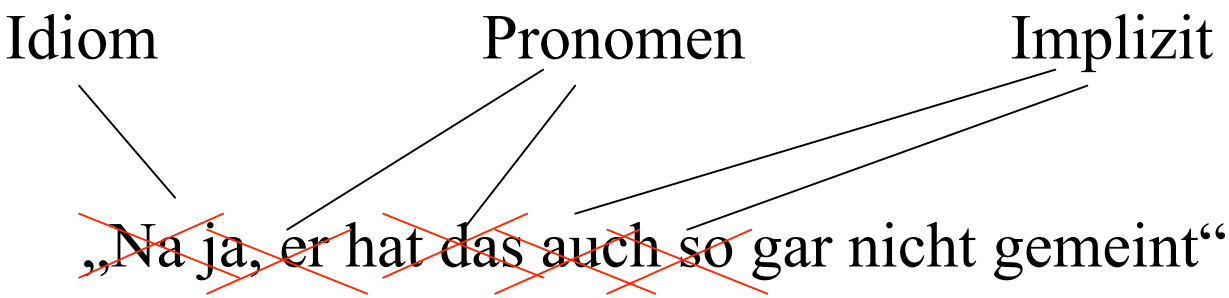
Text-Standards

- Lexikalische Regeln
  - Eingeschränkte Liste von Akronymen und Abkürzungen,
  - Eingeschränkte Liste von Synonymen,
  - Demonstrativpronomen sollen häufig benutzt werden.
- Syntaktische Regeln:
  - Kongruenz zwischen Subjekt und Prädikat muß genauestens beachtet werden
  - Ellipsen dürfen nicht benutzt werden
  - Regeln über Länge und Einbettungstiefe von NPs

Beispiele für häufige Regeln in kontrollierten Sprachen - 2 -

Text-Standards

- Textstruktur-Regeln:
  - Festlegungen, wann und wie Tabellen und Listen benutzt werden dürfen,
  - Eingeschränkte Anzahl von Sätzen in einem Paragraphen
- Pragmatische Regeln:
  - keine Idiome, keine Umgangsprache
  - Jede wichtige Information muß explizit sein (möglichst "ausbuchstabierte" Präsuppositionen)



Allgemeines zu Regeln in kontrollierten Sprachen -1-

Text-Standards

- Eine kontrollierte Sprache hat zwischen 30 und 60 Regeln und ein Lexikon.
- Überwiegend sind es lexikalische und syntaktische Regeln.
- Textuelle Regeln sind weniger häufig, wenn, dann in HOCL.
- Pragmatische Regeln sind sehr selten, weil sie sehr schwer formal definierbar und daher nur schwer durch einen automatischen Korrektheits-Checker überprüfbar sind.
- Nach O'Brien 2003 "An Analysis of Several Controlled Language Rule Sets", findet man nur 1 Regel in allen 8 untersuchten kontrollierten Sprachen :
- "Keep procedural sentences as short as possible (20-25 words) "

Regeln in kontrollierten Sprachen -2-

Text-Standards

- Regeln die man wenigstens in der Hälfte der untersuchten kontrollierten Sprachen findet:
  - "Use approved words from the dictionary"
  - "Make your instructions as specific as possible"
  - "When appropriate, use an article (the, a , an) or a demonstrative adjective (this, these) before a noun
  - "Use active voice"

Lesbarkeits- vs. Übersetzbarkeits-Regeln

Nach Reuther 2003 “Readability and Translatability by means of Controlled Language”):

- Lesbarkeit braucht weniger Regeln als Übersetzbarkeit
- Lesbarkeitsregeln sind eine Untermenge von Übersetzbarkeitsregeln
- Übersetzbarkeitsregeln sind spezifischer als Lesbarkeitsregeln
- Übersetzbarkeitsregeln sind teilweise eine Vorform von “Preediting”.
- Beide sind szenarioabhängig

Text-Standards

Übersetzbarkeitsregeln
- Beispiele-

- Formatierungsregeln (wegen Tokenisierungproblemen):
  - Wörter oder Satzteile eventuell durch Klammern strukturieren
- Grammatik-Regeln (wegen Parsing-Problemen)
  - kein Passiv
  - keine doppelte Negation
  - keine ambigen Genitiv-NPs

Die Highlands-Landschaft enthält Seen, Moor und Berge

Text-Standards

Verbesserung der Lesbarkeit in HOCL

Es ist eine wunderbare, schöne Landschaft

Es gibt aber wenig Menschen, die dort leben

- Even the glorious loneliness of the Highland's wonderful landscape of loch, moor and mountain is largely a product of the 'clearances' of the 18th and 19th centuries, which caused so much hardship and suffering

- The Highlands of Scotland consist of lakes mountains and moors. The moors are flat empty lands where no trees grow. The land is wonderful and magnificent because it is so empty. However, many people once lived here. But in the 18th and 19th centuries the owners of the land forced these people to leave. These people suffered many difficulties and troubles. People call these terrible events 'the Clearances'

Das ist wegen der 'Clearances'

'Clearaces" war eine Bewegung in 18ten und 19ten Jahrhundert.

Wegen der Clearnaces haben Menschen viel gelitten.

Controlled Language Authoring Technology (CLAT) - 1 -

- Werkzeuge, die technische Redakteure bei der Erstellung qualitativ hochwertiger Dokumentation unterstützt.
- Funktionen:
  - Rechtschreibprüfung,
  - Kontrolle von kundenspezifischer Terminologie und Abkürzungen,
  - Grammatikprüfung gemäß allgemeinsprachlichen und firmenspezifischen Regeln,
  - Stilkontrolle gemäß allgemeinen Richtlinien in der technischen Dokumentation und kundenspezifischen Schreibregeln (Stilregeln).

Controlled Language Authoring Technology (CLAT) -2-

- CLAT Sprachen:
  - Deutsch
  - Englisch
  - Für Französich, Italienisch, Spanisch, Schwedisch: Prototypen
- Plattformen: Sun, Linux, Windows NT, XP, 2000, Mac OSX
- Benutzungsfreundliche graphische Benutzungsoberfläche

Aus: http://www.iai.uni-sb.de/iaide/de/clat.htm

Text Standards

Lokalisierungsstandards

- The Localisation Industry Standards Association (LISA) is conducting an Industry Survey. Its questionnaire will be found at the following URL:

  <http://www.lisa.unige.ch/99survey_form.html>

- The LISA is looking for feedback from any developers and users of MT systems as well as localisation tools. The more answers the more authoritative and reliable the results of the survey for everyone interested in the field of MT.