Treebanks: Current Trends and Future Perspectives

Erhard W. Hinrichs
University of Tübingen
eh@sfs.uni-tuebingen.de

A Walk in the Woods

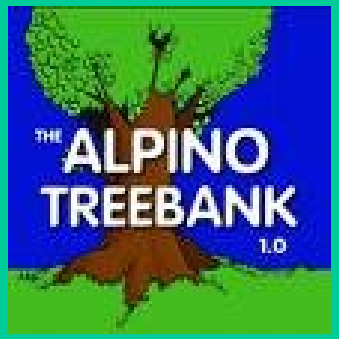
Erhard W. Hinrichs
University of Tübingen
eh@sfs.uni-tuebingen.de

A Macabre Joke

Baumbank

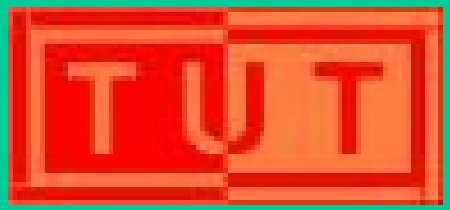
Goals of this Talk

- To reflect on more than 30 years of treebank development
- To reflect on five years of TLT workshops
- To look toward the future
- To stimulate discussion at this workshop


THE
ALPINO
TREEBANK
1.0


ÚFAL


S
T


TUT


GENiA


hobs


Arboretum

And there are many more ...

- METU-Sabanci Turkish Treebank
- Danish Dependency Treebank
- Penn Korean and Chinese Treebanks
- Penn Treebank of Middle English
- ...

Treebanks and Linguistic Theories

- CCG Bank
- Prague Dependency Treebank
- Redwoods Treebank (HPSG)
- Automatic LFG-Grammar Induction from treebanks
- Automated extraction of Tree-Adjoining Grammars from treebanks

Multi-stratal Annotation

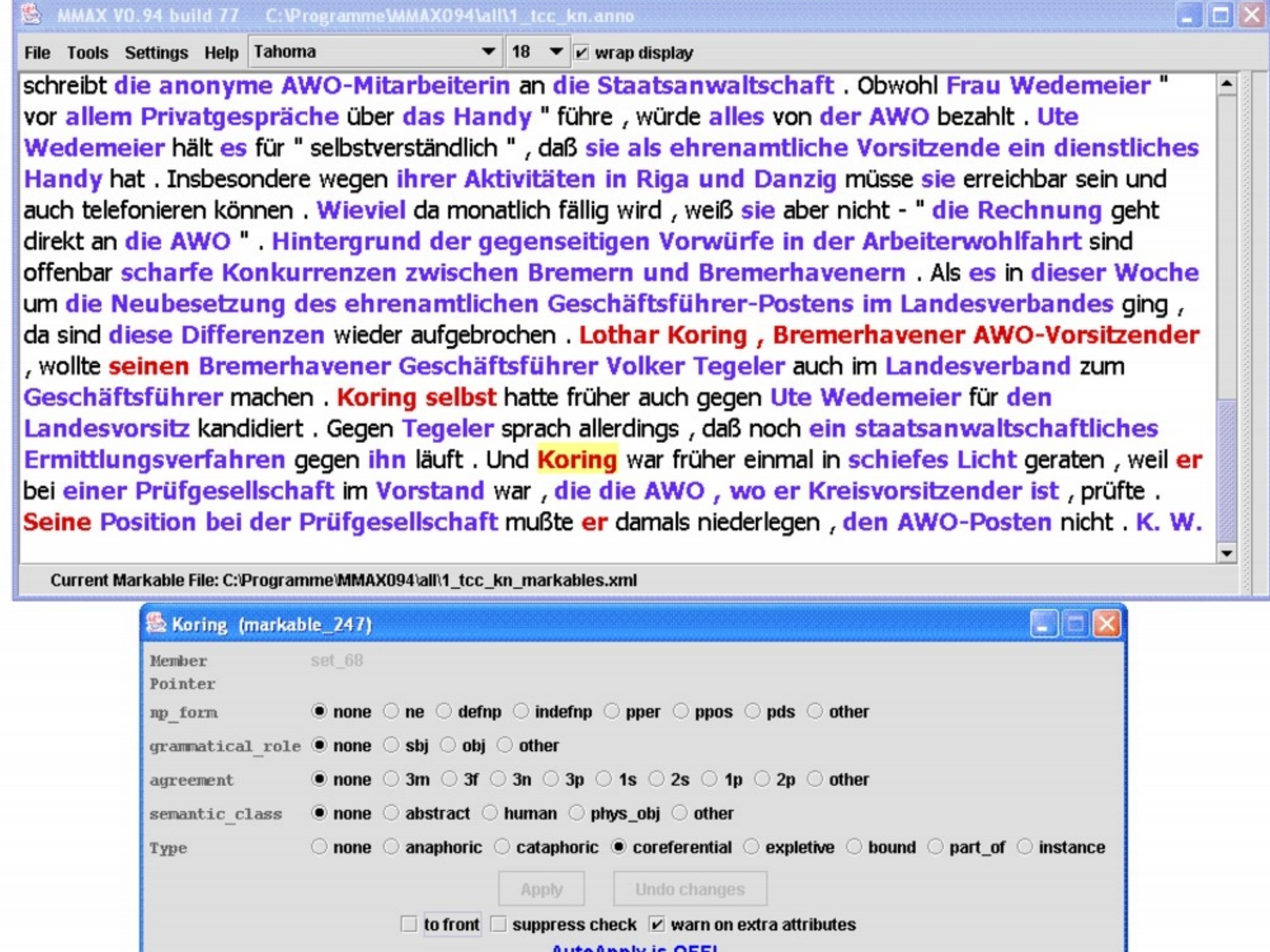
Three Mainstays

- Morphological/POS annotation
- Syntactic annotation
- Argument structure

Multi-stratal Annotation

Recent Arrivals:

- Co-reference annotation
- Semantic Annotation
- Temporal annotation
- Discourse annotation


MMAX V0.94 build 77 C:\Programme\MMAX094\all\1_tcc_kn.anno
File Tools Settings Help Tahoma 18 wrap display
schreibt die anonyme AWO-Mitarbeiterin an die Staatsanwaltschaft . Obwohl Frau Wedemeier " vor allem Privatgespräche über das Handy " führe , würde alles von der AWO bezahlt . Ute Wedemeier hält es für " selbstverständlich " , daß sie als ehrenamtliche Vorsitzende ein dienstliches Handy hat . Insbesondere wegen ihrer Aktivitäten in Riga und Danzig müsse sie erreichbar sein und auch telefonieren können . Wieviel da monatlich fällig wird , weiß sie aber nicht - " die Rechnung geht direkt an die AWO " . Hintergrund der gegenseitigen Vorwürfe in der Arbeiterwohlfahrt sind offenbar scharfe Konkurrenzen zwischen Bremern und Bremerhavenern . Als es in dieser Woche um die Neubesetzung des ehrenamtlichen Geschäftsführer-Postens im Landesverbandes ging , da sind diese Differenzen wieder aufgebrochen . Lothar Koring , Bremerhavener AWO-Vorsitzender , wollte seinen Bremerhavener Geschäftsführer Volker Tegeler auch im Landesverband zum Geschäftsführer machen . Koring selbst hatte früher auch gegen Ute Wedemeier für den Landesvorsitz kandidiert . Gegen Tegeler sprach allerdings , daß noch ein staatsanwaltschaftliches Ermittlungsverfahren gegen ihn läuft . Und Koring war früher einmal in schiefes Licht geraten , weil er bei einer Prüfgesellschaft im Vorstand war , die die AWO , wo er Kreisvorsitzender ist , prüfte . Seine Position bei der Prüfgesellschaft mußte er damals niederlegen , den AWO-Posten nicht . K. W.
Current Markable File: C:\Programme\MMAX094\all\1_tcc_kn_markables.xml
Koring (markable_247)
Member set_68
Pointer
np_form none ne defnp indefnp pper ppos pds other
grammatical_role none sbj obj other
agreement none 3m 3f 3n 3p 1s 2s 1p 2p other
semantic_class none abstract human phys_obj other
Type none anaphoric cataphoric coreferential expletive bound part_of instance
Apply Undo changes
to front suppress check warn on extra attributes

Challenge I: ULA

- Development of standard encodings for multistratal annotation
  - Annotation graphs/GrAF/GENAU
- Development of graphical browsing tools
  - Extending **annotate, TigerSearch, XBANK**
- Develop annotation tools for multistratal annotation

NLP Application Areas

- **Statistical Parsing**
- Machine Translation
- Computational Lexicography

Corpus Variation and Parser Performance (Gildea 2001)

Training Data	Test Set	Recall	Prec.
WSJ	WSJ	86.1	86.6
WSJ	Brown	80.3	81.0
Brown	Brown	83.6	84.6
WSJ+Brown	Brown	83.9	84.8
WSJ+Brown	WSJ	86.3	86.9

Interested in planting or replacing a tree on your treebank?

Citizens are required to obtain a permit from the City of Elgin Department of Public Works prior to planting a tree in the treebank area of your property. Treebank trees should only be trimmed or removed by the City of Elgin. To obtain a permit or request that a tree be trimmed or removed, please call (847)931-6069.

Challenge II

How to Balance your Treebank:

- Tree Entropy (Hwa)
- Creating Language Samplers (BNC Sampler)
- Profiling Lexical Coverage

The Limits of Manual Annotation

„Many uses of structurally annotated corpus data require a scale that is unrealistic to achieve by manual annotation.“

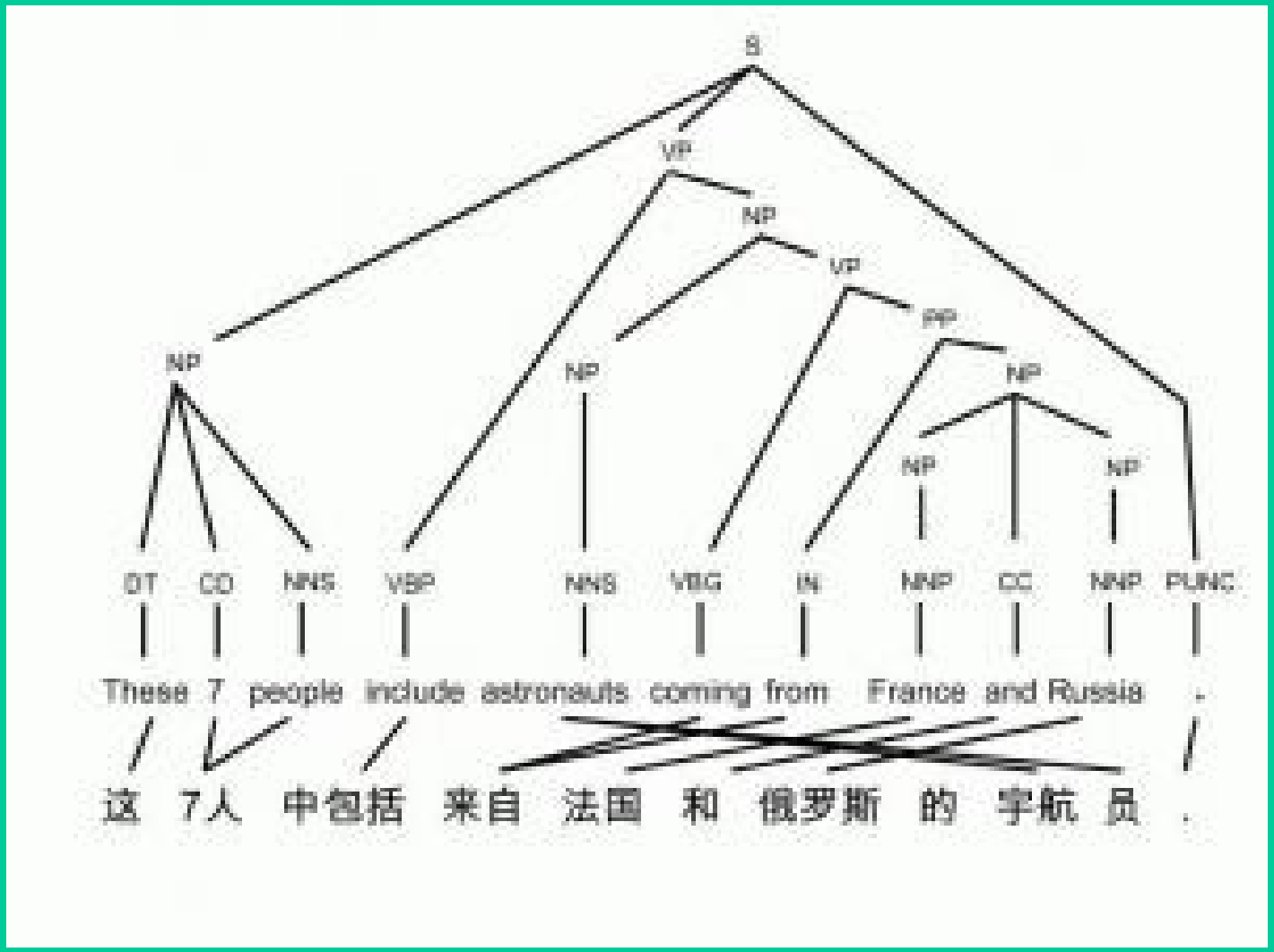
(Jonas Kuhn, ULA workshop, Bergen)

NLP Application Areas

- Statistical Parsing
- **Machine Translation**
- Computational Lexicography

Parallel Treebanks

- Prague Czech-English Dependency Treebank
- Swedish Parallel Treebank
- Nordic Parallel Treebank


S
VP
NP
VP
PP
NP
NP
NP
NP
NP
DT
CD
NNS
VBP
NNS
VBG
IN
NNP
CC
NNP
PUNC
These 7 people include astronauts coming from France and Russia .
这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

Knight and Marcu

"We propose to implement a trainable tree-based language model and parser, and to carry out empirical machine-translation experiments with them. USC/ISI's state-of-the-art machine translation system already has the ability to produce, for any input sentence, a list of 25,000 candidate English outputs. This list can be manipulated in a post-processing step. We will re-rank these lists of candidate string translations with our tree-based language model, and we plan for better translations to rise to the top of the list."

NLP Application Areas

- Statistical Parsing
- Machine Translation
- **Computational Lexicography**

Extracting Selectional Preferences

by Latent Semantic Clustering (Rooth 1998)

- Method for the extraction of selectional preferences of verbs from large quantities of data for the resolution of attachment ambiguities
- Verbs with similar selectional preferences and their preferred objects end up in the same cluster
- Soft clustering: Elements may be member of more than one cluster

Cluster Extracted from TüBa-D/Z

Verbs		Objects	
geben *'give'*	0.90958	Alternative *'alternative'*	0.01918
starten *'start'*	0.02361	Antwort *'answer'*	0.01476
ankündigen *'announce'*	0.01623	Mühe *'effort'*	0.01033
unterrichten *'teach'*	0.00738	Auskunft *'information'*	0.01033
plazieren *'place'*	0.00464	Meinung *'opinion'*	0.00885
überreichen *'hand over'*	0.00443	Position *'position'*	0.00754
aktivieren *'activate'*	0.00443	Absprache *'agreement'*	0.00738
durchspielen *'run through'*	0.00318	Möglichkeit *'possibility'*	0.00738
		Licht *'light'*	0.00738
Vernachlässigen *'neglect'*	0.00295	Krieg *'war'*	0.00600
leihen *'lend/borrow'*	0.00212		

Automatic Syntactic Annotation

TüPP/D-Z

- automatically annotated using a cascaded finite state parser chunk analysis
- grammatical functions added in a second pass
- about 11.5 million sentences (appr. 204.6 million tokens)

Cluster Extracted from TüPP-D/Z

Verbs		Objects	
sagen *'say'*	0.04944	Menschen *'people'*	0.036425
verletzen *'injure'*	0.029765	Frau *'woman'*	0.013469
töten *'kill'*	0.024556	Mann *'man'*	0.012581
glauben *'believe'*	0.017252	Leute *'people'*	0.012347
erschießen *'shoot'*	0.013988	Kinder *'children'*	0.011219
fragen *'ask'*	0.013367	Frauen *'women'*	0.011097
meinen *'believe'*	0.010223	Personen *'persons'*	0.007007
ermorden *'murder'*	0.009509	Männer *'men'*	0.006798
angreifen *'attack'*	0.009457	Soldaten *'soldiers'*	0.005445
festnehmen *'arrest'*	0.007927	Opfer *'victim'*	0.004726

Challenge III

- How to create very large automatically annotated corpora
- This requires
  - robust parsers with high accuracy
  - good error detection tools

How to find the smelly trees

Opportunities

- Strengthening the ties between research on annotated corpora and linguistic theories
- Community Building

CLARIN
Common Language Resources and Technology Infrastructure

CLARIN:

A pan-European Research Infrastructure Initiative for Language Resources and Technology

- has about 90 member institutions from 31 European countries
  - includes most of the well-known researchers and technologists from our field
  - has commitment statements from 25 member states (growing)

  - two tier organizational structure
    - European level with EC support
    - national networks of CLARIN members (national coordinator)

- total EC budget support of 4.1 Mio € for 3 years prep phase

SIGANN

ACL SPECIAL INTEREST GROUP

- The exchange and propagation of research results with respect to the annotation, manipulation and exploitation of annotated language resources, taking into account different applications and theoretical investigations in the field of language technology and research;
- Working towards harmonization and interoperability of linguistic annotations from the perspective of the increasingly large number of tools and frameworks tht support the creation, instantiation, manipulation, and exploitation of annotated resources;
- Working towards consensus on all issues critical to the advancement of the field of language resource annotation.

The Future of TLT

- ✓TLT 2008: Groningen (co-located with CLIN)

Expression of Interest: Milan, Sofia

On the wishlist: Dublin, Tartu

Thank you for your attention!