PROCEEDINGS OF SCIENCE

metaDictionary – Towards a Generic e–Infrastructure for Detecting Variance in Language by Exploiting Dictionary Information

Dietmar Seipel* [a] **and Werner Wegstein** [b]
University of Würzburg, Am Hubland, D – 97074 Würzburg, Germany
[a] *Department of Computer Science*
[b] *Institute for German Philology*
E–mail: {dietmar.seipel,werner.wegstein}@uni-wuerzburg.de

In this paper, we investigate the problem of building a metaDictionary that reflects variance in a language in space and time. This is done by exploiting the annotation of electronic dictionaries; for obtaining a fine–grained annotation, we use declarative parsing techniques.

We are developing a graphical tool for decomposing and annotating the entries/morphemes of an electronic dictionary. The goals of the segmentation are to obtain a list of basic morphemes, to construct a network showing which morphemes can be combined to obtain complex morphemes, and to analyze variance in space and time. Based on the metaDictionary and the annotated morpheme decomposition, a network analysis of morpheme decompositions becomes possible, which can take variance and grammatical properties into account.

We can manage and analyze grids of dictionaries and huge text corpora: dictionaries are especially interesting for our analyses, since they contain settled morpheme cores, and the text corpora document which morphemes can be combined lexicographically. Our dictionary–based morphological analysis has to be tested in the context of huge network corpus data best accessible via grid structures.

The International Symposium on Grids and Clouds and the Open Grid Forum
March 19–25, 2011
Academia Sinica, Taipei, Taiwan

*Speaker.

PoS(ISGC 2011 & OGF 31)003

Contents

1. Introduction

Information and knowledge can be encoded by combining elementary basic components according to special rules. From this point of view, genomes and language code may have structural properties in common which contribute substantially to biological evolution on the one hand and to the change of language on the other hand. Research in this field with a focus on interdependencies between science and the humanities seems promising. In bioinformatics, research data on genomes have already been generated and are publicly available. On a much smaller scale, we try to build up a comparable research environment in the humanities by collecting empirical data in the field of language change in German within the last 500 years and beyond, based on dictionary information and to be used for the research into mutual relations between data structures and data properties in genomes and language. Our interdisciplinary project combines bioinformatics, informatics, philology and corpus linguistics and has been funded by the German Federal Ministry of Education and Research since 2008 [12].

The starting point of the project is a collection of digitized dictionaries made publicly available by our project partners at the university of Trier (*Trierer Wörterbuchnetz* [11]): synchronic dictionaries like the Middle High German Dictionaries (Benecke/Müller/Zarncke and Lexer), early High German Dictionaries around 1800 (Adelung and Campe), a selection of dictionaries on regional dialects (Rheinisches, Pfälzisches and Luxemburger Wörterbuch, Wörterbuch der elsässischen bzw. der deutsch–lothringischen Mundarten) and supporting materials of the diachronic *Deutsches Wörterbuch* by Jacob and Wilhelm Grimm. For modern German, we can use Klappenbach/Steinitz, *Wörterbuch der deutschen Gegenwartssprache (WDG)*, digitized by the Academy of Sciences of Berlin and Brandenburg [4]. Based on the broad and representative data base, the goal is to develop and test methods and algorithms for detecting and understanding variance. Ideally, variance in biological base structures could then be modelled using concepts detected by analyzing language structures, and, vice versa, variance in language might be described using models developed for structures in bioinformatics.

The rest of this paper is organized as follows: In Section 2, we outline some project objectives investigating variance in language and genome. Section 3 explains a prerequisite of the project: how digitized print dictionaries can be annotated in TEI using a declarative grammar formalism. Section 4 presents a tool for analyzing the morphological structure of dictionary entries and annotating their parts. Some conclusions are given in Section 5.

2. Variance in Language and Genome

Our project goals are the compilation of a metaDictionary – using a fine–grained annotation of dictionaries – and the computation of network structures and properties for the comparison with genomes – based on morphological analyses of entries into basic lexical units (base morphemes). As early as 1934, Karl Bühler discussed the importance of these units for the understanding of variety in language in his language theory and gave an estimate on the dimensions for German – more than 2000 units –, setting out with a morpheme count of 30 pages of Goethe's novel *Wahlverwandtschaften* [1]. In this section, we will briefly sketch the project goals and the techniques from computer science that we are using; more details about the techniques will be given in the subsequent sections.

2.1 The metaDictionary

The usage of words in languages varies in space and time, cf. Figure 1. A metaDictionary summarizes different representations of the same lexical units, taken from the dictionaries spread in space and time, in lists of metaLemmas, cf. Figure 2. The metaLemma is based on dictionary entries in present–day standard German, with special provision made for units that are no longer used today.

2.2 Network Analysis of Morpheme Decompositions

The project partner from bioinformatics is interested in comparing the combinability of (basic) morphemes in words, cf. Figure 3 with the combinability of amino sequences

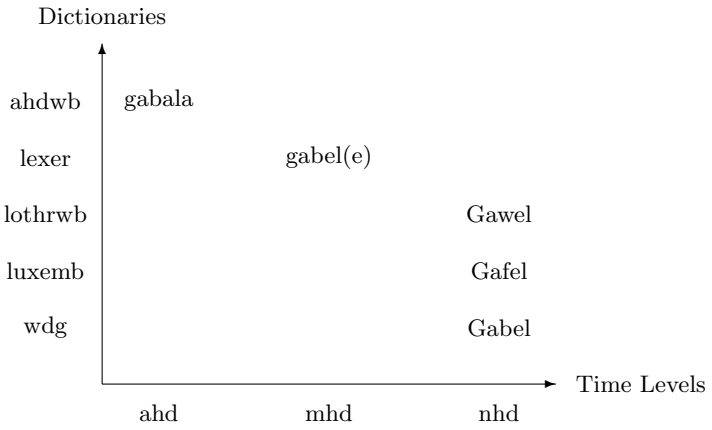


Figure 1: Variance in Space and Time.

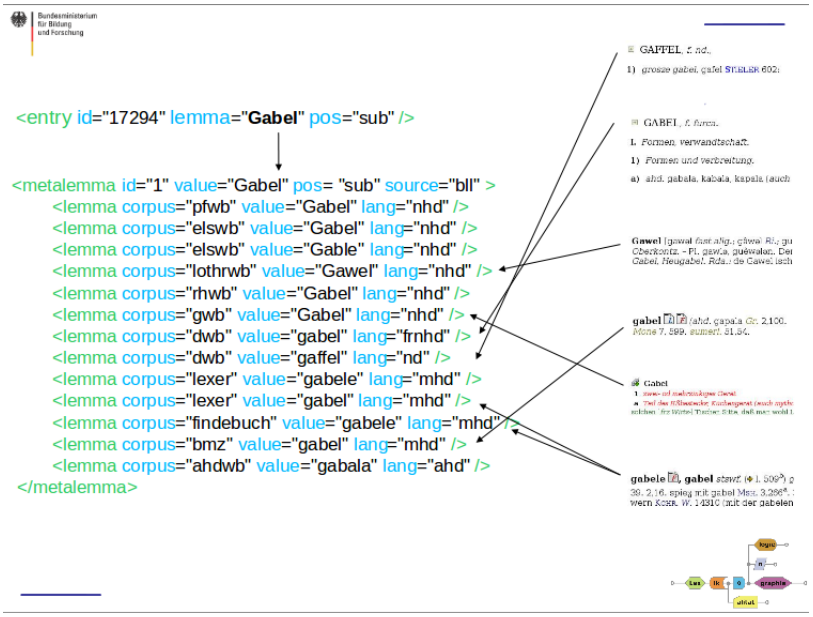


Figure 2: A metaLemma Represents a List of Dictionary Entries.

in genomes. Preliminary tests have shown that the corresponding networks have similar properties. This could be due to the fact that the generative processes behind the evolution of language and genome might be the comparable.

2.3 Techniques from Computer Science

We want to develop a generic e–Infrastructure for analyzing the inhomogenious dictionary entries of 18th, 19th and 20th century lexicographers and for transforming the information to a kind of baseline encoding keeping as much of the valuable dictionary information as possible, e.g., part of speech, gender, and inflectional detail, encoded in lots

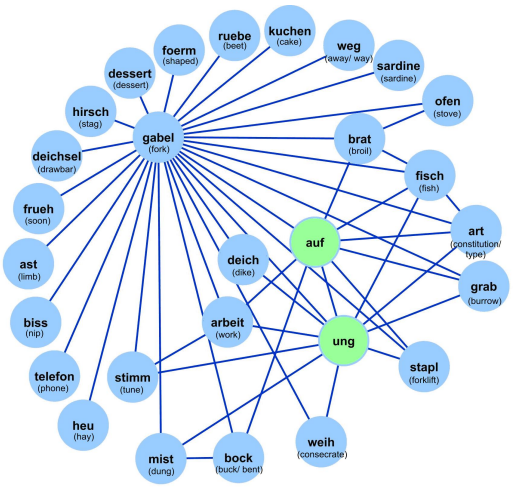


Figure 3: Morpheme Network.

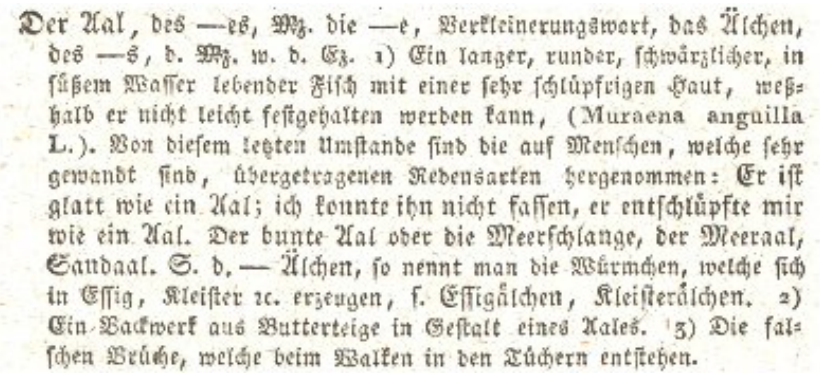
of different patterns. The annotated dictionary data are stored in an XML data format conformant with the TEI P5 Guidelines for Electronic Text Encoding and Interchange [9] for researching into variance comparable to genome structures, and they will be publicly available for other researchers.

We use *declarative programming* in PROLOG for parsing, querying, and transforming the linguistic data [3, 7, 8]. There exists quite a variety of well–known declarative languages for different areas: SQL for relational databases, XQuery and XSLT for XML processing, PROLOG for declarative programming, and rules for decision support systems and grammars. Their advantages are, that they are compact, rapidly programmable, clear, less error–prone, and flexibly extensible. We have tested declarative knowledge extraction by parsing dictionary entries with definite clause grammars (DCG) on the Adelung Dictionary and – within the framework of the TextGrid–Project [10] – on the Campe Dictionary [2] for a fine–grained lexicographic analysis. For compiling the metaDictionary, we have split the lexemes from the electronic dictionaries into morpheme terms. We have built a PROLOG–based tool for a declarative, rule–based control and annotation of morpheme segmentations, which we test on WDG data at present. We can extend our approach by alignment methods, which can yield a more precise alignment for the metaDictionary.

3. Annotation of Digitized Print Dictionaries in TEI

Researching morpheme structures requires a fine–grained analysis and detailed accounting of the basic linguistic units of the German vocabulary. Using dictionary entries as the basis helps to minimize influences of inflectional specialties and the variance of morphological combination patterns, irrespective of the permanent change of language in history. The dictionaries we selected fully represent the lexicographical state of the German language during the Middle High German Period (Lexer, supported by the Grimm dictionary) and New High German, used by educated and literary circles from around 1775

(Adelung, Campe) up to sixties in the 20th century (WDG) [13]. And we can apply the results of our dictionary analyses in a second step to text corpora of middle High German texts and early new High German texts – starting with Luther and the mass of German literary texts – available soon in the TextGrid digital repository [10].



Der Aal, des —es, Mz. die —e, Verkleinerungswort, das Älchen, des —s, d. Mz. w. d. Ez. 1) Ein langer, runder, schwärzlicher, in süßem Wasser lebender Fisch mit einer sehr schlüpfrigen Haut, weßhalb er nicht leicht festgehalten werden kann, (Muraena anguilla L.). Von diesem letzten Umstande sind die auf Menschen, welche sehr gewandt sind, übergetragenen Redensarten hergenommen: Er ist glatt wie ein Aal; ich konnte ihn nicht fassen, er entschlüpfte mir wie ein Aal. Der bunte Aal oder die Meerschlange, der Meeraal, Sandaal. S. d. — Älchen, so nennt man die Würmchen, welche sich in Essig, Kleister ꝛc. erzeugen, s. Essigälchen, Kleisterälchen. 2) Ein Backwerk aus Butterteige in Gestalt eines Aales. 3) Die falschen Brüche, welche beim Walken in den Tüchern entstehen.

Figure 4: Entry of an Electronic Dictionary.

For automatically extracting the keyword (lexeme) and some meta–data, such as word class and inflexion, we use a compact grammar formalism, which we have developed using the declarative programming language Prolog. Parsing of dictionary entries is a complex task, since there is a lot of structural variance; if the desired data could not be extracted, then we can flexibly and quickly adapt the grammar rules without breaking other cases. The parsed dictionary data are stored in a Tei data format for researching into variance comparable to genome structures.

3.1 Encoding Dictionaries in Tei

With our parsing approach, we can obtain a fine–grained annotation of electronic dictionaries in a valid Tei format. E.g., the dictionary entry for *Aal* from Figure 4 is structured as shown in Figure 5:

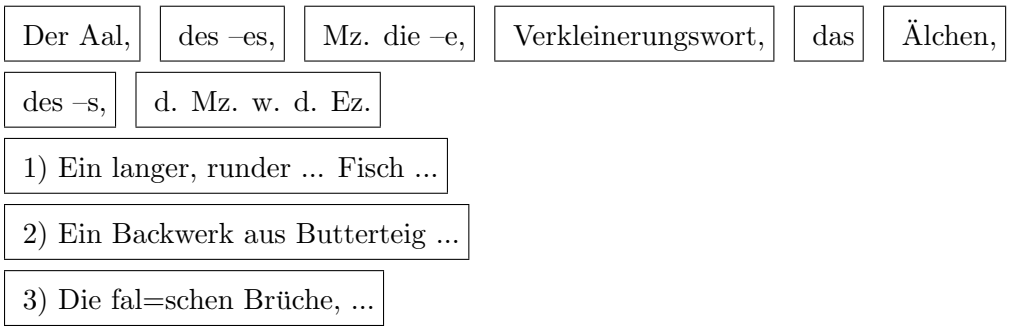


Figure 5: Fine–Grained Structuring.

The obtained structures are represented using Tei elements. E.g., the dictionary entry for *Aal* is annotated as follows:

```
<entry xml:id="cwds1_00005_aal">
```

```
   <form type="lemma">
      <gramGrp> <pos value="noun"/> <gen value="m"/> </gramGrp>
      <form type="determiner">Der</form>
      <form type="headword">Aal</form>
      <pc>,</pc>
   </form> ...
   <sense> ... </sense>
</entry>
```

XML is a common data format for modelling, managing, and exchanging semi–structured data. Powerful query, transformation and update languages exist for XML. We can also gain performance compared to relational databases, since XML is more appropriate for handling complex structures.

3.2 Grammar–Based Parsing

Parsing with *grammars* yields a higher precision compared to regular expressions and statistical parsers. We use a DCG extension, called Extended Definite Clause Grammar (EDCG) rules, which is even more compact and directly, generically generates XML [7]. E.g., for generating the entry elements, we can use the following EDCG rules:

```
entry ===>
   form:[type:lemma], ..., sense.
form:[type:lemma] ===>
   sequence(*, form:[type:determiner]),
   form:[type:headword].
sense ===> ...
```

The attribute `xml:id` of the `entry` element, which depends on the position of the entry in the dictionary (and on the headword), is determined in a further processing step. EDCGs offer meta–predicates such as sequence: the call `sequence(*, form:[type:determiner])` generates a sequence of zero or more `form` elements.

The `*`–notion is well–known from extended Backus–Naur form (EBNF) of context free grammars. Compared to DCGs and EBNF, an important extension of EDCGs is, that they generically generate XML. When working with DCGs in PROLOG or EBNF in other tools, this generation has to be coded explicitly and thus blows up the code drastically.

4. Annotating Morpheme Decompositions

For annotating the large numbers of dictionary entries (which can exceed 100.000 units), one needs linguistic knowledge and suitable tools from computer science: a reliable morphological analyzer, a suitable, compact knowledge representation, inference methods for automatic annotation rules, and a graphical user interface. The architecture of our annotation tool, which we have built using PROLOG, is shown in Figure 6.

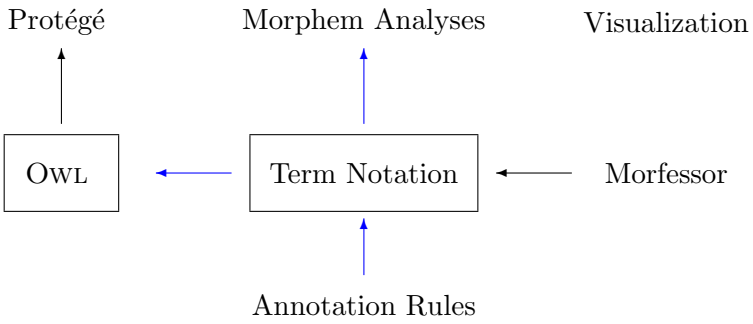


Figure 6: System Architecture.

The decomposition of complex morphemes into basic units is based on the Whole Word Morphology. The initial morpheme decompositions are derived using well–established tools such as Morfessor. Then, they are pre–annotated based on the fine–grained information of electronic dictionaries. We use dictionary information provided by the WDG to support the analysis, e.g., information on word classes and conditions of usage. The basic morphological units found can later be corrected, refined and annotated with our tool. We try to extract as much of the annotations as possible from the underlying electronic dictionaries using our parsing techniques, and we have developed a set of about 50 generic annotation rules for inferring further annotations. The final morpheme decompositions can be exported to the Web Ontology Language Owl and imported into standard tools such as Protégé [6].

4.1 Annotated Morpheme Terms

We use a compact knowledge representation of the annotated morpheme decompositions as suitable term structures that can be managed and extended elegantly in Prolog.

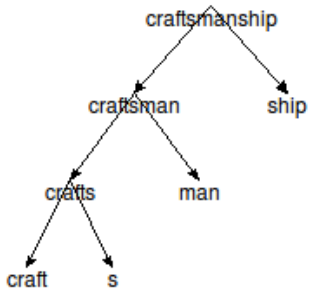


Figure 7: Morpheme Decomposition.

Using unique abbreviations, such as `bm` for basic morphem and `ge` for gap element, the morpheme term `((craft + s) + man) + ship` of the English word `craftsmanship`, which is visualized in Figure 7, can be annotated as follows:

```
((craft*bm + s*ge) + man)*noun + ship
```

For persistently storing the morpheme terms and for efficiently accessing the morpheme terms based on their text form, a relational database is used. For every morpheme term, it contains further information, such as the text form, the identifier from the WDG, and the name of the user who annotated the term as well as the timestamp of the annotation.

4.2 Annotation Rules

The processing of the morpheme terms is completely done on the PROLOG side. In the background, a set of PROLOG rules is managed, which can automatically infer further annotations from the annotations and decompositions given by the user. With the following logical annotation rule, the term `((craft + s) + man)*noun + ship` is recognized as a noun and can be further annotated to `(((craft + s) + man)*noun + ship)*noun`:

```
has_word_class(X, noun) :-
   mc(X, A, B), has_word_class(A, noun), text_form(B, [ship, ...]).
```

4.3 The Morpheme Annotation Tool

The morpheme annotation tool supports the user in checking all pre–annotated morpheme decompositions. The user can annotate parts of a morpheme decomposition. Based on a set of annotation rules – which can be extended dynamically – the morpheme decomposition is further annotated as detailed as possible. The graphical user interface, which is also implemented in PROLOG, consists of the morpheme editor, a visualization of the morpheme terms as tree structures, and a database browser, cf. Figure 8.

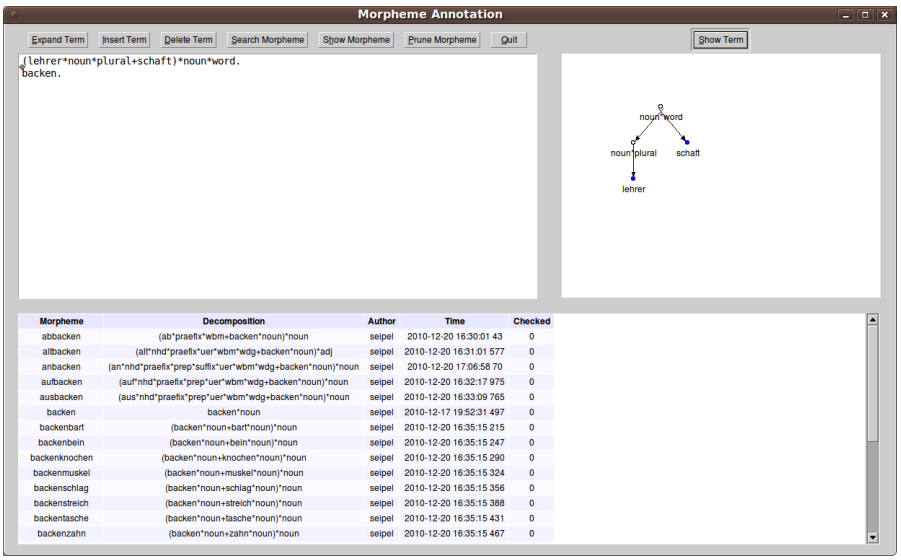


Figure 8: The Graphical User Interface of the Morpheme Annotation Tool.

The tree structure of the annotated morpheme terms can be visualized for better readability. The annotated morpheme terms are stored in a relational database. It is possible to efficiently search for individual annotated morpheme terms or for all morpheme terms containing a search string.

5. Conclusions

The metaDictionary of the German language, based on the analyses of a network of dictionaries, forms the core part of the generic e–Infrastructure designed to identify the basic morphemes used in the German language within the last 500 years. The next step will be to test the data using text corpora: first to check for basic morphemes not covered by dictionaries, and second to find out all combinations of basic morphemes used in texts. Here we expect new insights into the combinability of basic units, quantitatively as well as qualitatively, because dictionaries of the German language normally do not register complex morpheme structures representing semantically regular patterns. We will compare our results with the observations in *Culturomics*, that *52% of the English lexicon – the majority of the words used in English books – consists of lexical dark matter undocumented in standard references* [5]. We expect to contribute with our e–Infrastructure in a Grid environment to the development of test methods and algorithms for detecting and understanding variance in language that ideally are applicable to genome structures as well.

References

[1] Bühler, Karl: *Sprachtheorie. Die Darstellungsfunktion der Sprache.* Jena, 1934, p. 34.

[2] Campe, Joachim Heinrich: *Wörterbuch der deutschen Sprache.* 5 Vol., 1807–1811.

[3] Gazdar, Gerald; Mellish, Chris: *Natural Language Processing in* Prolog. *An Introduction to Computational Linguistics.* Addison–Wesley, 1989.

[4] Klappenbach, Ruth; Steinitz, Wolfgang (Eds.): *Wörterbuch der deutschen Gegenwartssprache.* 6 Vol., Akademie–Verlag, Berlin, 1961–1977.

[5] Michel, Jean–Baptiste, et al.: *Quantitative Analysis of Culture Using Millions of Digitized Books.* Science 14, January 2011, pp. 176–182.

[6] *The Protégé Ontology Editor.* `http://protege.stanford.edu/`.

[7] Schneiker, Christian; Seipel, Dietmar; Wegstein, Werner; Prätor, Klaus: *Declarative Parsing and Annotation of Electronic Dictionaries.* Proc. 6th International Workshop on Natural Language Processing and Cognitive Science (NLPCS), 2009.

[8] Seipel, Dietmar: *Processing* Xml*–Documents in* Prolog. Proc. 17th Workshop on Logic Programmierung (WLP), 2002.

[9] *P5: Guidelines for Electronic Text Encoding and Interchange.* `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html`.

[10] *Textgrid: A Modular Platform for Collaborative Textual Editing – a Community Grid for the Humanities.* 2009, `http://www.textgrid.de`.

[11] *Trierer Wörterbuchnetz – Network of Electronic Dictionaries of the University of Trier.* `http://www.woerterbuchnetz.de/`.

[12] *Variance in Language and Genome*: Project funded by the German Ministry of Education and Research (BMBF) since 2008. `http://www.sprache-und-genome.de`.

[13] Gouws, Rufus et. al.: *Wörterbücher / Dictionaries / Dictionnaires. ... / An International Encyclopedia of Lexicography / ... Vol. 2*, de Gruyter, Berlin / New York, 1990.