Morphological Paradigms: Computational Structure and Unsupervised Learning

Jackson L. Lee
University of Chicago
jsllee@uchicago.edu

Abstract

This thesis explores the computational structure of morphological paradigms from the perspective of unsupervised learning. Three topics are studied: (i) stem identification, (ii) paradigmatic similarity, and (iii) paradigm induction. All the three topics progress in terms of the scope of data in question. The first and second topics explore structure when morphological paradigms are given, first within a paradigm and then across paradigms. The third topic asks where morphological paradigms come from in the first place, and explores strategies of paradigm induction from child-directed speech. This research is of interest to linguists and natural language processing researchers, for both theoretical questions and applied areas.

1 Introduction

Morphological paradigms (e.g., *walk-walks-walked-walking*) are of central interest to both linguists and natural language processing researchers for the connectedness (e.g., *jumps, jumping* sharing the lexeme JUMP) and predictability across words (e.g., inducing *googles* for *google* based on *jump-jumps* etc). This thesis explores the computational structure of morphological paradigms, particularly from the perspective of unsupervised learning for modeling how such structure can be induced from unstructured data. Three topics under study are as follows:

- **Stem identification:** The first part of the thesis concerns the structure *within* a morphological paradigm, focusing on stem identification. The goal is to devise general and language-independent strategies for stem extraction applicable for different types of morphology across languages, and goes beyond the common substring-based approaches.

- **Paradigmatic similarity:** The second part of the thesis asks what structure there is *across* morphological paradigms. Paradigms often do not inflect in the exact same pattern, which leads to inflection classes, e.g., Spanish verbs in distinct conjugation groups. At the same time, paradigms inflect in remarkably similar ways, e.g., Spanish verbs in the second plural all end with *-mos* regardless the inflection classes. This part of the thesis develops a string-based hierarchical clustering algorithm that computationally characterizes the similarity and differences across morphological paradigms.

- **Induction of morphological paradigms from unstructured data**: The third part of the thesis seeks to induce paradigms from unstructured data. The kind of unstructured data of interest here is child-directed speech. Building on previous work on unsupervised learning of morphological paradigms from raw text, this thesis develops an approach of paradigm induction that incorporates results from the previous two parts of this thesis and has a version taking child-directed speech data incrementally.

These three topics on morphological paradigms progress in terms of the scope of data in question. The first and second parts explore structure when paradigms are given – one paradigm at a time, and then a list of paradigms together. The third part asks where morphological paradigms come from in the first place. This research will be of interest to both linguistics (the nature of strings, morphemes, and paradigms) and natural language processing (information retrieval, machine translation).

2 Stem identification

Given a morphological paradigm with inflected word forms, what is the stem of the paradigm? This question on stem identification is part of the morpheme segmentation problem, important for both theoretical linguistics (Spencer 2012) and computational linguistics (Goldsmith 2010, Hammarström and Borin 2011); once the stem is identified, what is not the stem in each word form can be subject to further segmentation and morphological analysis for potential affixes. Stem identification is far from being a trivial problem. Strictly concatenative morphology, as exemplified by English *jump-jumps-jumped-jumping* with "jump" as the stem, appears intuitively simple. In contrast, non-concatenative morphology, a well-known case being Arabic root-and-pattern morphology (e.g., *kataba* 'he wrote', *yaktubu* 'he writes/will write' with "k-t-b" as the stem) has been treated as something fundamentally different. The first part of this thesis seeks to develop language-independent, algorithmic approaches to stem identification which are sufficiently general to work with both concatenative and non-concatenative morphology.

2.1 Linearity and contiguity

The problem of stem identification begins with the definition of "stem" in a morphological paradigm. A common and language-independent assumption is that the stem (broadly construed, encompassing "root" and "base") is the maximal common material across all word forms in the paradigm. This thesis explores different definitions of "maximal common material" in search of general algorithms of stem identification for languages of different morphological types. In particular, we examine ways of characterizing strings in terms of linearity and contiguity.

As a point of departure, we take the maximal common material to mean the maximal common *substring*, a very intuitive and common assumption in morpheme segmentation. To illustrate the idea of a substring with respect to linearity and contiguity, consider the string "abcde". "a", "bc", and "cde" are its substrings. "ac" is not a possible substring, because "a" and "c" are not contiguous. "ba" is not a substring either, because "a" does not linearly come after "b" in the string "abcde". Because substrings embody both linearity and contiguity, if a stem in a morphological paradigm is the longest common substring across the word forms, then this approach of stem identification works well only for strictly concatenative morphology but not for anything that deviates from it. To solve this problem, this thesis explores various ways of defining the maximal common material with regard to linearity and contiguity.

2.2 Substrings, multisets, and subsequences

The definition of maximal common material may depend on whether linearity and contiguity are respected. Three major definitions along these two parameters are of interest; see Table 1:

	Substring	Multiset	Subsequence
Linearity	✓	✗	✓
Contiguity	✓	✗	✗

Table 1: Three definitions of maximal common material for stem identification in terms of linearity and contiguity

(The possibility of maintaining contiguity but abandoning linearity results in pairs of symbols which appear to be less informative for stem identification.)

As noted above, defining the stem as the maximal common *substring* is suboptimal for non-concatenative morphology. The two other strategies consider the stem as the maximal common *multiset* or *subsequence*, illustrated in Table 2 by the Spanish verb PODER 'to be able' conjugated in present indicative. Taking the stem to be the maximal common *multiset* yields the set $\{p,d,e\}$ as the stem for the PODER paradigm. Table 2 highlights the stem material for each word form. Certain word forms have multiple stem analyses because of the multiple occurrences of "e" in the words concerned; these can be resolved by cross-paradigmatic comparison in section 3 below or paradigm-internal heuristics (e.g., choosing the stem that is the most congruent with non-stem material compared to other words in the paradigm, as in Ahlberg et al. 2014). In contrast, if the stem is the maximal common *subsequence*, then there are two competing stems for the PODER paradigm: p-d and p-e (using '-' to denote linear order without committing to contiguity). These two stems are tied because they each contain two symbols and are the longest possible common subsequences in the paradigms.

	Multiset	Subsequence	
	{p,d,e}	p-d	p-e
puedo	**p**u**ed**o	**p**ue**d**o	**p**u**e**do
puedes	**p**u**ed**es **p**ue**de**s	**p**ue**d**es	**p**u**e**des **p**ued**e**s
puede	**p**u**ed**e **p**ue**de**	**p**ue**d**e	**p**u**e**de **p**ued**e**
podemos	**p**o**de**mos	**p**o**d**emos	**p**od**e**mos
podéis	**p**o**de**is	**p**o**d**eis	**p**od**e**is
pueden	**p**u**ed**en **p**ue**de**n	**p**ue**d**en	**p**u**e**den **p**ued**e**n

Table 2: Stem as maximal common multiset or subsequence for the Spanish PODER paradigm conjugated for present indicative

The subsequence approach has clear merits. Recent work—both directly and indirectly on stem identification—appears to converge on the use of the subsequence approach (Fullwood and O'Donnell 2013, Ahlberg et al. 2014). This is because it can handle Arabic-type non-concatenative morphology, infixation, circumfixation (as in German *ge*-X-*t*), and (trivially) the *jump*-type strictly concatenative morphology. In general, linearity appears to be more important than contiguity in stem identification. It must be noted, however, that probably for the more familiar properties of substrings, linguists are accustomed to using multi-tier substrings to handle surface non-contiguity, e.g., McCarthy (1985) on templatic morphology and Heinz and Lai (2013) on vowel harmony.

This part of the thesis serves as the foundational work for the later parts. For this first part, languages of interest include those with morphology diverging from simple concatenation, e.g., English with weak suppletion, Spanish with stem allomorphy, Arabic with templatic morphology, and German with circumfixation. Datasets come from standard sources such as Wiktionary (cf. Durrett and DeNero 2013). In terms of evaluation, a particular stem identification algorithm can be tested for whether it provides the correct stems for paradigm generation, an evaluation method connected to the clustering of paradigms in section 3.

Apart from stems, stem identification necessarily identifies the residual, non-stem material in each word form in the paradigm. The non-stem material is analogous to the affixes and stem allomorphs (e.g., the *o*∼*ue* alternation in PODER). It plays an important role in terms of structure across morphological paradigms, the subject of the next section.

3 Paradigmatic similarity

The second part of the thesis asks what structure there is *across* morphological paradigms. Word forms across paradigms do not alternate in the same pattern. Linguists discuss this in terms of inflection classes, which introduce differences across morphological paradigms. At the same time, however, morphological patterns are also systematically similar. This part of the thesis focuses on the modeling of *paradigm similarity* and develops a string-based hierarchical clustering algorithm that computationally characterizes the similarity and differences across morphological paradigms, with both theoretical and practical values.

3.1 Inflection classes

Morphological paradigms often do not inflect in the same way, which leads to inflection classes. For example, Spanish verbs are classified into three conjugation groups (commonly referred to as -AR, -ER, and -IR verbs), illustrated in Table 3 for the inflectional suffixes (all person and number combinations) in present indicative.

	-AR	-ER	-IR
1.SG	-o	-o	-o
2.SG	-as	-es	-es
3.SG	-a	-e	-e
1.PL	-amos	-emos	-imos
2.PL	-áis	-éis	-ís
3.PL	-an	-en	-en

Table 3: Suffixes for the three Spanish conjugation groups in present indicative

The Spanish conjugation classes show what is common across languages that this part of the thesis models: *partial* similarity across morphological paradigms. Spanish is described as having three conjugation classes for the three distinct overall suffixing patterns. For example, they are completely different for first-person plurals (*-amos, -emos*, and *-imos*). At the same time, they share a great deal in common. Across all three classes, the first-person singular suffixes are *-o*, the second-person singular suffixes end with *-s*, and so forth. Some classes share properties to the exclusion of others: the second and third conjugation groups share *-es, -e, -en* for 2.SG, 3.SG, 3.PL respectively, but the first conjugation group have *-as, -a, -an* instead.

The similarities and differences which morphological paradigms exhibit as inflection classes are of interest to both linguistics and natural language processing. In linguistics, the partial similarities across inflection classes prompt theoretical questions on the extent to which paradigms can differ from one another (Carstairs 1987, Müller 2007). Computationally, inflection classes introduce non-uniformity across paradigms and must be handled in one way or another in an automatic morphology learning system. Previous work has opted to explicitly learn inflection classes (Goldsmith and O'Brien 2006) or collapse them in some way (Chan 2006, Hammarström 2009, Monson 2009, Zeman 2009).

3.2 Clustering for paradigm similarity

This thesis aims to characterize paradigm similarity in a way that is amenable to a linguistic analysis and a formal model of paradigm similarity useful for computational tasks related to paradigms. As discussed above, similarities and differences criss-cross one another in morphological paradigms and result in inflection classes. It is therefore reasonable to think of morphological paradigms as having a string-based hierarchical structure, where paradigms more similar to one another by the inflectional patterns cluster together. Haspelmath and Sims (2010) explore just this idea using data from Greek nouns and demonstrate how inflection classes can be modeled as a problem of clustering, though their work appears to be based purely on the human linguist's intuition and is not computationally implemented. This thesis proposes a string-based hierarchical clustering algorithm (with morphological paradigms as the objects of interest to cluster) for modeling paradigm similarity, which is (i) built on results of stem identification from section 2 and (ii) useful for further computational tasks such as paradigm generation.

There are multiple advantages of proposing a clustering algorithm for morphological paradigms. To the linguist, results of clustering paradigms can be visualized, which will be helpful for the study of inflectional structure of the morphology of less familiar languages (such as those based on fieldwork data). For computational linguistics and natural language processing, clustering provides a similarity measure that is useful for inducing unobserved word forms of incomplete morphological paradigms.

The proposed algorithm performs agglomerative hierarchical clustering on a given list of morphological paradigms. It involves stem identification (section 2) that determines the non-stem material in the word forms of each paradigm. The distance metric measures similarity among the paradigms by comparing non-stem material, which forms the basis of the distance matrix for hierarchical clustering.

Preliminary work (Lee 2014) suggests that clustering morphological paradigms gives desirable results. To illustrate, Figure 1 shows the clustering results of our algorithm under development for several English verbal paradigms (by orthography). For reasons of space, the results of only ten English verbs are discussed here; see Lee (2014) for details.

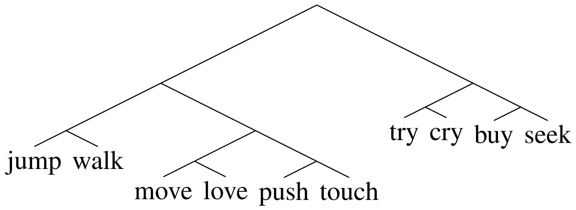

Figure 1: Simplified clustering results for a few English verbal paradigms, each represented by the infinitive form

In Figure 1, the two largest clusters of verbs are the one with more regular morphology on the left (JUMP, WALK, MOVE, LOVE, PUSH, TOUCH) and the other on the right with verbs of more drastic inflectional/orthographic alternations (TRY, CRY with the *i*∼*y* alternation, and BUY, SEEK with *-ght* in past tense). The smaller clusters among the regular verbs are due to the form for third person singular in present tense (PUSH, TOUCH with an additional 'e') and the verb-final 'e' (MOVE, LOVE with 'e' dropped for the *-ing* form). This example shows that clustering morphological paradigms provides a much more fine-grained characterization of inflection classes, which are usually described in non-hierarchical terms in linguistics.

An open question here is how to evaluate the results of clustering morphological paradigms. The major issue is that morphological paradigms are usually not hierarchically represented in standard descriptions, thereby making it unclear what the gold standard could be. One possibility is that the learned inflection classes (based on clusters of paradigms) be compared to those in standard grammatical descriptions of the language in question. Alternatively, the results can be evaluated indirectly by what the induced structure should facilitate, namely paradigm generation; this also evaluates stem identification in section 2. Datasets of paradigm tables for languages with inflection classes (English, Greek, Spanish, etc) come from standard sources such as Wiktionary. Paradigm generation takes a paradigm table with held-out words for some paradigms, and the goal is to recover the missing words using (i) stems computed based on the available words in the respective paradigms (section 2) and (ii) non-stem material as predicted based on the cross-paradigmatic cluster information (this section).

4 Induction of morphological paradigms from unstructured data

The discussion so far has assumed that a list of morphological paradigms are available for the study of structure within (section 2) and across (section 3) paradigms. While this is a common practice in the cognitive and computational modeling of morphological paradigms (Albright and Hayes 2002, Durrett and DeNero 2013), it is legitimate to ask where a list of morphological paradigms come from in the first place. This part of the thesis attempts to provide an answer to this question. Building on previous work on unsupervised paradigm induction, this thesis will propose a language-independent, incremental paradigm learning system that induces paradigms with child-directed speech data as the input.

4.1 Incremental paradigm induction

The unsupervised learning of morphological paradigms has attracted a lot of interest in computational linguistics and natural language processing (Goldsmith 2001, Schone and Jurafsky 2001, Chan 2006, Creutz and Lagus 2005, Monson 2009, Dreyer and Eisner 2011, Ahlberg et al. 2014). Virtually all previous work proposes a batch algorithm of paradigm induction, rather than an online and incremental learner, that takes some raw text as the input data. This is probably cognitively implausible, because a human child does not have access to all input data at once. This thesis proposes an incremental paradigm induction system to fill this gap of the relative lack of work on the incremental and unsupervised learning of morphological paradigms.

As a starting point, the proposed paradigm induction system will use one akin to Linguistica (Goldsmith 2001) and adapt it as an incremental version. The choice of a system like Linguistica as the point of departure is justified, because the goal here is to induce morphological paradigms from unstructured data but not necessarily morpheme segmentation (accomplished by other systems such as Morfessor (Creutz and Lagus 2005) that focus strongly on morphologically rich languages such as Finnish and Turkish). Linguistica induces paradigms by finding the optimal cut between a stem and an affix across words that could enter into paradigmatic relations, and does not perform further morpheme segmentation. A characteristic of Linguistica that will be modified in this thesis is that of stem identification: as it currently stands, it assumes (i) strictly concatenative morphology (i.e., stem as maximal common *substring*), and (ii) knowledge of whether the language under investigation is suffixing or prefixing. In line with the general goal of coming up with language-independent algorithms to handle natural language morphology, we will make use of the results from section 2 on stem identification for languages of diverse morphological types.

The input data will child-directed speech from CHILDES (MacWhinney 2000) for North American English. Specifically, we will be using a dataset of four million word tokens compiled from child-directed speech data of age range from a few months old to 12 years old. The proposed algorithm will make use of the temporal information of the child-directed speech and read the data in small and chronologically ordered chunks. As such, this incremental version of Linguistica models child language acquisition, and the results will be of much interest to linguists. For evaluation, research on the child acquisition of English morphology (Cazden 1968, Brown 1973) provides the gold standard information on the order of acquisition of major morphological patterns (plurals acquired before possessives, present progressives acquired before pasts, etc).

4.2 Collapsing paradigms of different inflection classes

A recurrent problem in unsupervised learning of morphological paradigms is that certain induced morphological paradigmatic patterns may appear incomplete (due to unobserved word forms) or distinct on the surface (due to inflection classes), but should intuitively be collapsed in some way (Goldsmith 2009). For inflection classes, for instance, English verbs display a regular morphological pattern as in *Ø-s-ed-ing* (e.g., for JUMP), but there is also a very similar—but distinct—pattern, with *e-es-ed-ing* (e.g., for MOVE with the silent 'e'); this English example is by orthography, but is analogous to Spanish verbs with inflection classes discussed above. Ideally, it would be desirable to collapse morphological patterns, e.g., the two English morphological patterns just mentioned as belonging to the verbal category and with the correct morphosyntactic alignment for the suffixes across the two patterns. Previous work either ignores this issue and treats the distinct surface patterns as is (e.g., Goldsmith 2001) or attempts to collapse morphological patterns (e.g., Chan 2006, with the assumption of part-of-speech tags being available).

This thesis will explore the possibility of collapsing paradigms of different inflection classes with no annotations (e.g., part-of-speech tags) in the input data. Some sort of syntactic information will have to be induced and combined with the induced morphological knowledge, in the spirit of previous work such as Higgins (2002) and Clark (2003). We are currently using graph-theoretical approaches to the unsupervised learning of syntactic categories. Based on Goldsmith and Wang's (2012) proposal of the word manifold, a given corpus is modeled as a graph, where the nodes are the words and the edges connect words that are distributionally similar based on n-grams from the corpus. The resulting graph has distributionally (and therefore syntactically) similar words densely connected together, e.g., modal verbs and infinitives in Figure 2. Various graph clustering algorithms are being explored for the purposes of word category induction.

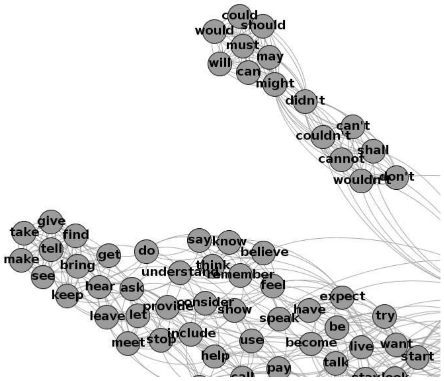

Figure 2: A zoomed-in image of clusters of modal verbs and infinitives in a 1,000-word graph

5 Contributions

This thesis will contribute to both the unsupervised learning of natural language morphology as well as bringing theoretical linguistics and computational linguistics closer together.

On the unsupervised learning of natural language morphology, this thesis explores structure within and across morphological paradigms and proposes algorithms for adducing such structure given a list of morphological paradigms. Furthermore, we also ask how an unsupervised learning system can induce morphological paradigms from child-directed speech, an area much less researched than previous work on non-incremental and batch algorithms for paradigm induction.

As for bridging theoretical linguistics and computational linguistics, this thesis represents a serious attempt to do linguistics that is theoretically informed from the linguist's perspective and is computationally rigorous for implementation. Using natural language morphology as an example, this thesis shows the value of reproducible, accessible, and extensible research from the computational community that will benefit theoretical linguistics.

References

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 569-578. Gothenburg, Sweden.

Adam Albright and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the 6th meeting of the ACL Special Interest Group in Computational Phonology*.

Roger Brown. 1973. *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Andrew Carstairs. 1987. *Allomorphy in Inflexion*. London: Croom Helm.

Courtney Cazden. 1968. The acquisition of noun and verb inflections. *Child Development* 39: 433-448.

Erwin Chan. 2006. Learning probabilistic paradigms for morphology in a latent class model. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL 2006*, 69-78. New York City.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics* (volume 1), 59-66.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 106-113.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from palin text using a dirichlet process mixture model. In *Proceedings of Empirical Methods in Natural Language Processing*, 616-627.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Michelle A. Fullwood and Timothy J. O'Donnell. 2013. Learning Non-concatenative Morphology. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (CMCL).

John A. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2): 153-198.

John A. Goldsmith. 2009. Morphological analogy: Only a beginning. In James P. Blevins and Juliette Blevins (eds.), *Analogy in Grammar: Form and Acquisition*, 138-164. Oxford: Oxford University Press.

John A. Goldsmith. 2010. Segmentation and morphology. In Alexander Clark, Chris Fox, and Shalom Lappin (eds.), *Handbook of Computational Linguistics and Natural Language Processing*, 364-393. Oxford: Wiley-Blackwell.

John A. Goldsmith and Jeremy O'Brien. 2006. Learning inflectional classes. *Language Learning and Development* 2(4): 219-250.

John A. Goldsmith and Xiuli Wang. 2012. Word manifolds. University of Chicago, ms.

Harald Hammarström. 2009. Unsupervised Learning of Morphology and the Languages of the World. Ph.D. thesis, Chalmers University of Technology and University of Gothenburg.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37(2): 309-350.

Jeffrey Heinz and Regine Lai. 2013. Vowel Harmony and Subsequentiality. In Andras Kornai and Marco Kuhlmann (eds.) *Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13)*, 52-63.

Jackson L. Lee. 2014. Automatic morphological alignment and clustering. Technical report TR-2014-07, Department of Computer Science, University of Chicago.

Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*. London: Hodder Education, 2nd ed.

Derrick Higgins. 2002. A Multi-modular Approach to Model Selection in Statistical NLP. University of Chicago Ph.D. thesis.

Brian MacWhinney. 2000. *The CHILDES Project*. New Jersey: Lawrence Erlbaum Associates.

John J. McCarthy. 1985. Formal Problems in Semitic Phonology and Morphology. New York: Garland.

Christian Monson. 2009. ParaMor: From Paradigm Structure to Natural Language Morphology Induction. Ph.D. thesis, Carnegie Mellon University.

Gereon Müller. 2007. Notes on paradigm economy. *Morphology* 17: 1-38.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1-9. Association for Computational Linguistics.

Andrew Spencer. 2012. Identifying stems. *Word Structure* 5(1): 88-108.

Daniel Zeman. 2009. Using unsupervised paradigm acquisition for prefixes. In *Evaluating Systems for Multilingual and Multimodal Information Access*, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarthus, Denmark, September 17-19, 2008, Revised Selected Papers, 983-990. Springer-Verlag, Berlin.