

Optimality in Analysis, Generation, and Learning: Towards a Robust Computational Architecture for Corpus-based Studies of Syntax¹

Jonas Kuhn

The University of Texas at Austin, Department of Linguistics /
Universität des Saarlandes, Saarbrücken, Germany

jonask@coli.uni-sb.de

This paper describes a computational architecture for accessing implicit information about the grammar of the languages included in a parallel corpus and exploiting it in an Optimality Theory-style learning approach. Previous work on OT learning presupposes the existence of training data in which the underlying input has been annotated. This is an idealization that does not reflect the natural learning situation; and it also requires considerable effort to produce such training data for learning experiments with syntactic/semantic grammar models. In the proposed bootstrapping architecture, which will be underlying in the new PTOLEMAIOS project, the training data are sentences from a parallel corpus; manual annotations are only provided for a small set of seed sentences. The translations of the sentence into the other languages serve as clues for zeroing in on the assumed underlying meaning representations, which can be used as the input in OT-style learning.

1 Introduction: the goal for the PTOLEMAIOS project

In this programmatic paper, an architecture for grammar learning based on parallel corpora is outlined. This proposed architecture is the target for the PTOLEMAIOS project.² More concretely, the project goal is to develop a formal architecture and implement a software system that allows one to train a syntactic grammar for a language L from a parallel corpus including L and multiple other languages, for which a relatively small set of sentences has been hand-annotated for syntactic correspondences across the languages. The resulting grammar should be robust and have broad coverage, while

¹ I would like to thank the audience at the KNAW Colloquium for valuable comments. Parts of this article overlap with Kuhn 2005a.

² PTOLEMAIOS is for “Parallel Text-based Optimization for Language Learning—Exploiting Multilingual Alignment for the Induction of Syntactic Grammars.” The project is funded in the Emmy Noether program (phase II) of DFG (German Research Foundation) as a research group led by the author at Saarland University in Saarbrücken. The project’s starting date was April 2005.

generally providing reliable analyses at the level of (at least clause-local) head-dependent relations; this will make the grammar applicable in natural language processing (NLP) applications that involve syntactic/semantic parsing at moderate depth, such as (multilingual) information extraction, question answering, or advanced statistical machine translation.

We can also phrase the project goal as a methodological challenge: to develop a formal framework for grammar learning which is sophisticated enough to allow for the integration of insights and assumptions from linguistic theory, and at the same time surface-oriented, robust, and computationally efficient enough so it can be applied on large amounts of real corpus data, without presupposing time-intensive manual annotation of more than a small subset of the data. We expect that with this methodological goal, our project results will transcend the immediate engineering achievements and contribute to our general understanding of the learnability of linguistic knowledge. In particular, our software architecture will serve as an empirical test bed for linguistic representation systems (e.g., of lexical classes, functional/lexical category distinctions, morphological marking, argument structure, etc.) with respect to learnability properties—an aspect for which so far it has been very hard to test theoretical predictions empirically.

In section 2, we motivate the proposed system architecture from the point of view of Optimality Theory (OT); section 3 contains a more detailed description of the planned architecture and representations. Section 4 provides further background on the design decisions made and points out some expected advantages. We close with a conclusion in section 5.

2 Motivation from the Optimality Theory perspective

Previous experimental work on OT learning (e.g., Boersma and Hayes 2001, Bresnan et al. 2001) has been presupposing the existence of training data in which the underlying input has been annotated for each learning instance. Irrespective of the specific learning algorithm assumed, the learning system has to know the underlying input in order to determine the candidate set. The system also has to know the intended winner, so it can determine whether the winner that the system obtains with the current ranking is identical to the intended winner, or whether a constraint reranking has to occur. In expressive optimization, the intended winner is typically determined by the directly observable surface part of the analyses (although there may be several string-identical options); the underlying input on the other hand cannot be read off the observed surface data. Hence, a learning model presupposing input-output data for training is an idealization that does not reflect the natural learning situation. From a practical point of view, there is the additional disadvantage that for performing simulations with an input-output training approach, considerable effort is required for annotating the training data. Annotation may still be feasible for grammar models of major distinctions in clausal syntax, but when studying more subtle phenomena that involve interactions between various fine-grained lexical distinctions, the amount of annotated training data that would be required for statistically significant

experimental results is prohibitive. At the same time, our understanding of such subtle interactions would arguably benefit most from experimental work based on real corpus data.

Conceptually, the standard assumption about the learning situation in OT is that the learner is able to apply some *robust interpretive parsing* procedure to obtain the underlying input for a given observed surface form. However, to our knowledge there have not been any attempts to spell out such a procedure in the context of syntactic/semantic OT models. Impressionistically, it is clear that the human learner can exploit a combination of quite diverse knowledge sources in order to decide on the underlying input: Firstly, the grammar in its current state will only allow for a limited number of possible readings for the observed surface form (even if it is applied in a somewhat “relaxed” mode, in order to anticipate that the grammar still needs to be adjusted); however, in order to be able to learn the complex, highly ambiguous/underspecified system of a natural language grammar, there will still be a large number of possible readings that cannot be excluded on grammar-internal grounds. But in addition to the prior grammatical knowledge, the learner’s perception of the utterance situation may narrow down the space of possible interpretations (e.g., through the presence/absence of entities that could be referred to); lastly, general knowledge or assumptions about the world may make certain interpretations much less likely than others.

It is not realistically possible (in the near future) to model this complex interaction of grammatical knowledge, extra-linguistic perception, and general inferences from further knowledge sources—as it would be required for a close simulation of the real learning situation. Therefore we propose to approximate the various extra-linguistic cognitive sources (which are “synchronized” with the linguistic utterance, to a higher or lesser degree) by something we can easily obtain in large amounts: the translations of a sentence into a number of other languages, as recorded in parallel corpora. Taking clues from various translations together, the learning system may be able to zero in on the correct underlying interpretation for a surface form that the system uses as a learning datum—so the parallel languages are functionally rather similar to the additional information sources in the impressionistic picture of the learning situation. What the learner needs in order to be able to exploit the additional information are the word and phrase correspondences across languages (roughly corresponding to the human learner’s need to be able to identify in the real world situation the referents of the phrases in an utterance). It seems realistic to assume that knowledge about such correspondences can be bootstrapped fairly reliably from an initial word alignment as it is produced by state-of-the-art techniques from statistical machine translation (compare the pilot study in Kuhn 2004).

The PTOLEMAIOS project will develop the methodology and tools for the outlined approximation of the real learning situation. The “OT-style” grammar models in PTOLEMAIOS differ from OT grammar models proper in one point: instead of OT’s constraint ranking model based on strict dominance, a more general class of probabilistic models—the log-linear or Maximum Entropy models—will be assumed. As has been observed, e.g., by Goldwater and Johnson 2003 and Jäger 2004, log-linear models are closely related to OT constraint ranking models. By using log-linear models, we can take

advantage of existing implementations of training algorithms, and since a log-linear grammar for expressive optimization is interpreted as a conditional probability model (assigning a probability to a candidate analysis, given an underlying input form), we can easily integrate the language-specific grammars in a larger probabilistic architecture. Finally, the move to a model without OT's strict dominance assumption is also motivated by considerations that this assumption may not carry over to interpretive optimization models in a straightforward way (compare Kuhn 2003, ch. 5), hence a less constrained model is a more appropriate basis for a symmetrical bidirectional optimization architecture. It should be noted however that it will be easy to move back to proper OT models in the PTOLEMAIOS architecture.

3 The computational architecture

Let us call the system architecture we plan to achieve in the project the PTOLEMAIOS system. Figure 1 illustrates the architecture with a flowchart.

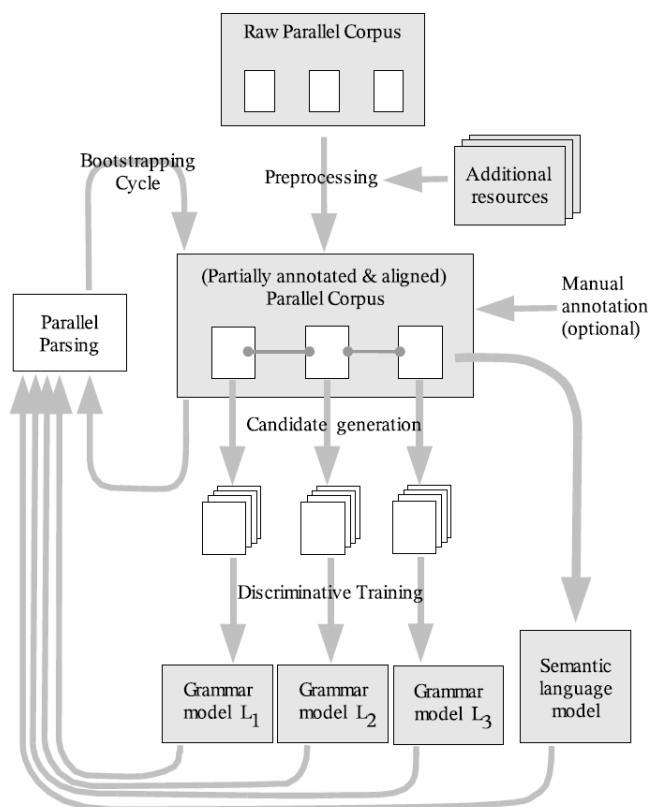


Figure 1: The PTOLEMAIOS architecture

The main input for the PTOLEMAIOS system is a parallel corpus, including translated text in at least two languages. As additional input, a subset of the parallel corpus is annotated with cross-linguistic information about phrasal correspondences and an underlying “pseudo meaning representation” which we will discuss in more detail below. We also use standard preprocessing techniques from NLP, such as part-of-speech tagging and morphological analysis, to the extent that the required resources are available. As additional preprocessing, (i) the parallel corpus is sentence-aligned, following the standard algorithm of Gale and Church (1991), and (ii), a statistical word alignment is trained using the GIZA++ tool (Och and Ney 2003).

Besides the preprocessing components, the PTOLEMAIOS system consists of a bootstrapping cycle for improving grammar models learned from the parallel corpus. The base learning component (section 3.1) at the core of the cycle involves candidate generation (creating generation alternatives to the one observed for each language in the corpus) and OT-style discriminative training of a grammar model. The outer loop (section 3.2) applies the grammar models of each stage to the corpus for the creation of a more accurate annotation as the basis for the next learning stage. The outer loop uses a noisy channel model for “parallel parsing”, inspired by the standard model underlying statistical machine translation.

3.1 Base learning in PTOLEMAIOS

The central grammar models obtained from the base learning component are log-linear expressive optimization models that predict for some language L , how likely it is to use a particular linguistic realization (let us call it t for “tree”) for a given underlying meaning representation m : we get a language-specific model for $P(t | m)$. The realization of a meaning with the highest probability can be viewed as our prediction of the correct grammatical realization of the meaning in language L .

Note that the model is based on an expressive optimization, i.e., the comparison of alternative realizations of the same underlying meaning (as it is assumed in most linguistic work in OT³), not an interpretive optimization as in classical statistical parsing where the task is to determine the most likely analysis tree for a given string. There are strong conceptual arguments for the expressive optimization architecture (see also section 4): the most crucial aspect in knowing the syntax of English can be paraphrased as knowing for example that one has to say *What did you see?*, rather than *What saw you?* or *Saw you what?* etc. In the initial implementation of the base learning component, we can build on an existing tool: the YASMET system, F. J. Och’s small toolkit for training conditional Maximum Entropy models.⁴ What needs to be implemented in the first phase of PTOLEMAIOS is a parser for the parallel corpus and routines for converting the parser output into the appropriate

³ For a discussion of directionality in formal OT syntax, see for instance Kuhn 2003.

⁴ <http://www.fjoch.com/YASMET.html>.

representation for the YASMET learning tool. The crucial conceptual points to clarify here are (i) what representation formats are used for the underlying meaning m and candidate analyses t , and (ii) how to parameterize the learning (in OT terms, what constraints to assume).

3.1.1 Representations

An important aspect of candidate analyses in OT is the distinction between a part defining the surface form and a part defining the underlying form (in syntactic OT, this typically corresponds to the meaning). For practical reasons, we use a comparatively simple candidate representation in the PTOLEMAIOS architecture: a candidate analysis is characterized by a tree in which the nonterminal symbols are augmented with atomic-valued features (from a finite set of features, each allowing for values from a finite set); let us call this representation format an augmented phrase structure tree (APT). There may be nonterminal symbols that do not dominate a terminal symbol; all terminal symbols are dominated by non-branching preterminal nodes. Consider for example the tree in figure 2, which we constructed for the English sentence *We will find out all the necessary information*, taken from the Europarl corpus (Koehn 2002). The EN-Nth and EN-Cat features encode the surface order of the daughters within the local subtree and the part-of-speech category of the syntactic head, respectively. Further feature distinctions for morphological form, sentence type etc. may be added. In our illustrations, only two categories are distinguished in the phrase structure backbone of the feature grammar: clausal nodes (CL) and non-clausal nodes (X).

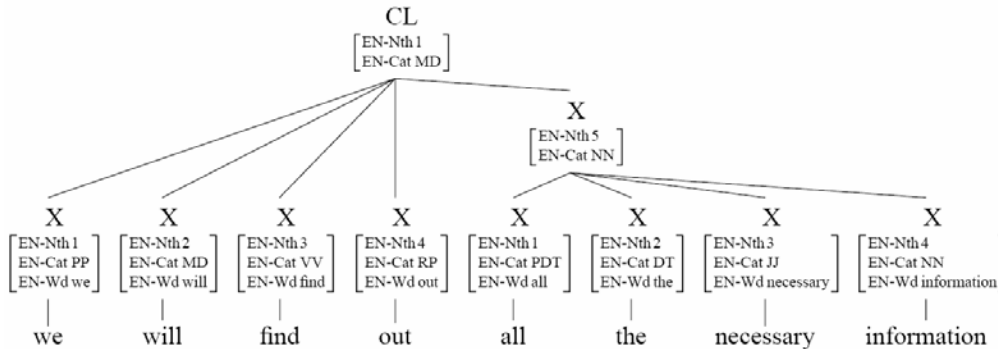


Figure 2: Augmented Phrase Structure Tree

The flat analysis of the clause-internal and NP-internal structure is motivated by the cross-linguistic context in which the analysis is used. We can assume very similar trees for translational correspondents as in (1) or the translations in (2).

- (1) EN We will find out all the necessary information
 FR Nous chercherons toutes les informations nécessaires
we will seek all the information necessary
 DE Wir werden alle notwendigen Informationen beschaffen
we will all necessary information acquire
- (2) EN I am not satisfied with what happened
 DE Nicht einverstanden bin ich jedoch mit dem was geschehen ist
not agreeing am I however with that what happened has

We go so far as to assume that translations have the same hierarchical graph structure—we may think of this as a consensus representation across languages. Formally, the multilingual consensus tree is generated by a variant of an inversion transduction grammar (ITG; compare Wu 1997). In the type of transduction grammar we assume, there are three ways in which the realization in a particular language may differ from the others: it is possible (i) to use different orderings of the daughters of a nonterminal, (ii) to use different nonterminal symbols for realizing a preterminal, and (iii) to leave certain nonterminal symbols unrealized.⁵

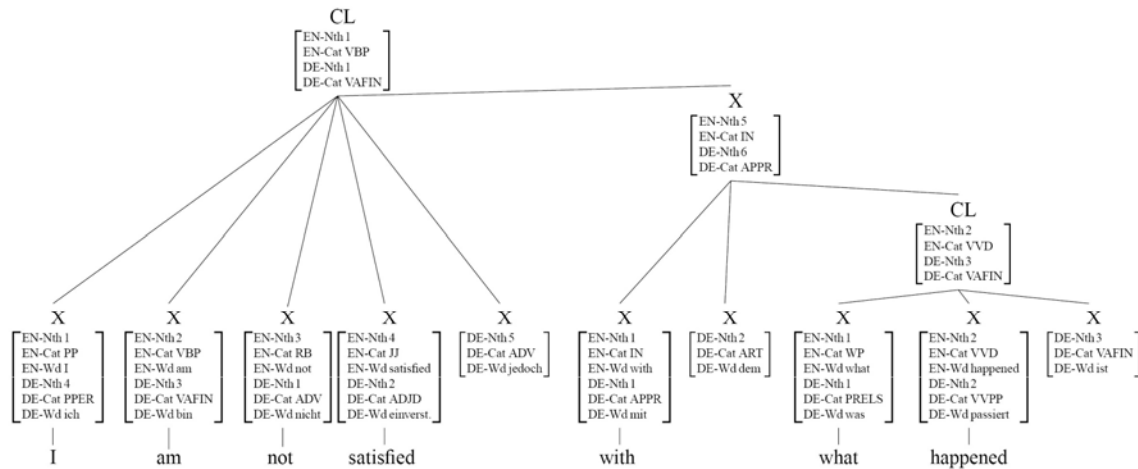


Figure 3: Candidate analysis for English, including pseudo meaning representation based on a German/English corpus

Given these assumptions we can draw a single graph for all translational realizations of a sentence. The graph for (2) shown in figure 3 includes the English and the German tree. (The actual terminal

⁵ A fourth option might be added at some point, using a special re-ordering operation for non-local realization (similar to the subtree cloning operation proposed by Gildea (2003)).

symbols and the sequencing for English is shown, but the reader may verify that the German version of (2) can be obtained by ordering the sisters in each local subtree according to the DE-Nth feature.) We can now specify the candidate analyses for PTOLEMAIOS. A candidate analysis is an augmented phrase structure tree (APT) like in figure 3, including information about language L and n other languages from a parallel corpus. The surface part of a candidate analysis for language L is simply the word string obtained by forming the sequence of the L -Wd feature values when traversing the tree according to the order of daughter nodes encoded in the L -Nth feature. The meaning part of a candidate analysis is essentially the consensus tree representation.

Due to the assumption of flat trees for entire clauses or argument phrases, most aspects of cross-linguistic syntactic variation can be construed as variable realizations of the underlying consensus tree, with differences in linearization, in category choice and choice of (non-)overt realization of certain nodes. This includes for instance basic word order, the choice of synthetic vs. analytic verb forms, *pro drop*, the use of expletives, placement of operator phrases, etc.⁶

How do we represent an underlying meaning independent of the particular realizations? The hierarchical structure is the most important aspect of the consensus tree representation; however, as a representation for syntactically relevant class distinctions of head items (e.g., *wh*-pronouns vs. personal pronouns, or different verb types), we also need labels for the nodes in the consensus tree. But since our approach is largely unsupervised and should be scalable, the class distinctions cannot be predefined and manually encoded. An additional issue is that to avoid a bias for a particular language, the labels should be language-neutral—coming from an interlingua, as it were. One possibility that we will explore is the use of “hidden” (or “latent”) class labels that are induced in a clustering approach; this means that an iterative process will form clusters among lexical categories based on their syntactic behavior as observed in grammar learning. But as a simple representational alternative to the choice of a single abstract label for each node, we will start out with the working hypothesis that the combination of category labels and word forms from *all* languages in the parallel corpus can be used as a complex “language-independent label” for a node. This means that the full APT which we proposed as a candidate representation is also our approximation of the underlying meaning. We will thus also call this APT representation (in which all terminal symbols are now ignored) the full “pseudo meaning representation” (full PMR).

A minor technical issue with the use of the fully merged language-specific representations is that if we want to train conditional grammar models $P(t_L | m)$ for a language L , using full PMRs as m , it will always be trivially the best choice to pick the L -realization that is included in m . So, for application in a conditional model, we will generally exclude the part of the full PMR on which the current decision

⁶ *In order to be able to express constraints that capture the syntactically relevant generalizations, a more fine-grained language-specific phrase structure (e.g., Grimshaw’s (1991) extended projections) has to be assumed on top of the flat consensus trees. This is fully compatible with the proposed architecture.*

is being made (or learned). We call this the relative PMR. The exact definition of relative PMRs depends on details of the conditional grammar model, but the crucial idea is that when a grammar for language L is learned, all L -specific features such as L -Wd and L -Nth are removed from the full PMR. That way, the learner can pick up systematic patterns in the translational correspondences. For instance, if French uses the pronoun *on* ('one') in an active sentence, while English uses a passive, it is likely to find either a passive or an active sentence with the pronoun *man* ('one') in German; whereas if French *on* and English *we* co-occur as subjects in the relative PMR for German, it is more likely to find an active sentence with the subject *wir* ('we') in German.

3.1.2 Parameterization of the expressive competition model

Our grammar models determine $P(t | m)$ for a particular language L . To estimate this conditional probability we will apply a log-linear or Maximum Entropy model (Ratnaparkhi 1999). This allows us to use a large number of OT-style constraints or learning features, which do not have to be statistically independent. Since many aspects of t are fixed even in the relative PMR format that we assume for m , only the following information needs to be determined by the model (relative to a specific language L): for a nonterminal node, whether or not it gets realized in L , and if it does get realized which is the syntactic head daughter, and what is relative order of the daughters. For preterminal nodes, the terminal realizing the node in language L has to be determined.

As mentioned above, most of the systematic cross-linguistic variation in the grammar of clauses and nominal phrases is reflected in ordering alternatives and the addition/omission of certain nodes, i.e., in the PMR these decisions are local to a subtree of depth 1. Therefore we will work with the following **separability assumption**: up to certain limited feature distinctions in our tree representations, the probability $P(t | m)$ for the entire tree can be separated out into a product of probabilities for local subtrees; these can be computed separately and (at least approximately) combined by dynamic programming techniques for unification grammars.⁷ This means that the candidate sets that have to be effectively computed and compared are comparatively small.

Since we plan to apply a log-linear model (rather than a simpler generative probability model), we can use fairly complex constraints in the training, reflecting linguistic insights into the levels of representation involved. Of course, we do not have access to a true, reliable semantic representation; it is a central hypothesis however that information from the other languages in the parallel text may effectively be used as a substitute. Thanks to the use of an expressive optimization we will be in a

⁷ This is related to locality observations for OT competitions based on extended projections, which was made in Kuhn 2001. Dynamic programming techniques for log-linear models of syntax are discussed by Geman and Johnson (2002).

position to experiment with the actual constraint sets assumed in the theoretical OT literature (to the extent that they can be adjusted to the specific representations we will assume).

3.2 The outer loop in PTOLEMAIOS

3.2.1 The bootstrapping cycle

Where does the PMR information (which is required to learn information about the grammar of language L) come from if no full manual annotation of the corpus is planned? The answer is that expressive optimization decisions are learned using a *bootstrapping approach* that involves several languages. Initially, either a small set of manually annotated sentence tuples from the parallel corpus is used as seed data (indicating clauses and major nominal/prepositional phrases and their syntactic heads), or an unsupervised grammar induction process is applied that exploits just the statistical word alignment (compare Kuhn 2004).

In each bootstrapping cycle, expressive-optimization-based log-linear grammar models are trained for each of the languages, making use of the relative PMR for the respective language and using only training data for which the full PMR has been previously assigned with high confidence. The result for each language L is a model for the conditional probability $P(t_L | m)$. Furthermore, the full PMRs are used to train an unconditional model for $P(m)$ (we can call this a “semantic language model”, following Miller et al. 1994), possibly using standard training of a probabilistic context-free grammar (PCFG) on a canonical form of the underlying PMR trees with some compiled-out features. The various models are then (re-)applied to analyze the full training data in “parallel parsing mode” (see below), trying to assign a PMR to the previously unanalyzed string tuples. Intuitively speaking, there are two options for each tuple of corresponding sentences: either all the grammars agree on a particular, common underlying PMR m . Or else, some of them prefer a meaning m_1 , while others suggest a different PMR m_2 , etc. In the former case, the combined system has high confidence in a particular consensus PMR, in the latter case, confidence is rather low. This confidence measure is exploited in order to decide on training data for the next bootstrapping step: only high-confidence sentence tuples are added to the pool of training data, and the consensus PMR m that was obtained for them is used in the next bootstrapping cycle to re-estimate the grammar models. In re-estimation based on a larger set of data, new patterns may be detected. So, as bootstrapping proceeds the log-linear grammar models can be trained on more and more “automatically annotated” data. Crosstalk between the grammars and the inherent redundancy of parallel corpora leads to an increase of the amount of usable training data (according to one of the central hypotheses of the PTOLEMAIOS project).

3.2.2 (Parallel) Parsing with a noisy channel model

Parsing tuples of translational correspondences from a parallel corpus simultaneously, using the current grammar models for the various languages, is the central procedure in the PTOLEMAIOS learning approach. It acts as an approximation of the human learner's combined use of cognitive systems and resources when learning from observed language data: when there are several possible analyses/underlying meanings of the observed linguistic utterance, visual information about the utterance situation, general knowledge about the type of situation, and various other non-linguistic information sources may be used in order to decide on what the intended meaning must be. For instance, *Put the doll into the house with the red roof* contains a PP attachment ambiguity. But if the learner can interpret the basic nominal phrases and sees that there are two houses, one with a red, the other with a blue roof, it is clear that the substring *the house with the red roof* has to form a unit that corresponds to a unit in the observed situation. Similarly for parallel corpus-based learning, the fact that the translational correspondents of all parts of a word sequence form a contiguous sequence (or not) may be exploited in order to draw statistical inferences about possible underlying consensus representations (compare Kuhn 2004).

The input to the parallel parsing step is (i) corresponding sentences in two or more languages from a parallel corpus, and (ii) a statistical word alignment for the sentence tuples. The *resources* used are the unconditional semantic language model and a log-linear grammar model for each of the languages under consideration. Let us call the models used a family of PMR transduction grammars. To keep the discussion simple, we will discuss parallel parsing for the case of parsing sentences in just two languages; the situation with more languages is analogous. The *output* of parallel parsing is the most probable full PMR underlying each pair (or tuple) of input sentences. The application of the various probability models follows the noisy channel model, which is underlying in all work on statistical machine translation. There, the idea is to solve the problem of translating, say, from French to English, not by training a direct model for translating a given French sentence into an English sentence (finding the English sentence e that maximizes $P(e|f)$). Instead, Bayes' law is used as in (3)⁸ to transform the problem into an equivalent combination of different subproblems, for which separate models are trained: (i) a *translation model* for determining how likely a French sentence is to be a translation of a given English source sentence (note that the assumed direction of translation underlying the training is reversed: a model for $P(f|e)$ is trained), and (ii) a translation-independent *language model* for determining how likely a sequence of words is as an English sentence ($P(e)$). To apply the two models in an actual translation task from French to English (!), a third component, a

⁸ The denominator can be ignored in the step from (3b) to (3c) since it remains constant when we are looking for the best e , given a fixed f .

decoder is needed, which searches the space of candidate word sequences e to maximize the product of the probabilities $P(f|e)P(e)$.⁹

$$\begin{aligned}
 (3) \quad & \text{a.} \quad \text{argmax}_e P(e|f) \\
 & \text{b.} \quad = \text{argmax}_e \frac{P(f|e)P(e)}{P(f)} \\
 & \text{c.} \quad = \text{argmax}_e P(f|e)P(e)
 \end{aligned}$$

Parsing can be viewed as a translation problem too: a sentence in a natural language is given, and the output is a translation of the sentence into some structural or semantic representation format. In statistical parsing, it is standard to approach this translation problem using direct estimation of the probability $P(\textit{tree} | \textit{sentence})$ or of the joint probability $P(\textit{tree}, \textit{sentence})$. In the PTOLEMAIOS project, we apply the noisy channel model to the parsing problem. In order to find the right tree for a given sentence, we search for the tree that maximizes the product $P(\textit{sentence} | \textit{tree})P(\textit{tree})$. The idea can be easily generalized to a scenario dealing with several languages (and using a pseudo meaning representation (PMR) m instead of the *tree*, as the common goal for parsing all corresponding sentences $s_{L1} \dots s_{Ln}$ —here we show the case for two languages).¹⁰

$$\begin{aligned}
 (4) \quad & \text{a.} \quad \text{argmax}_m P(m | s_{L1}, s_{L2}) \\
 & \text{b.} \quad = \text{argmax}_m \frac{P(s_{L1}, s_{L2} | m)P(m)}{P(s_{L1}, s_{L2})} \\
 & \text{c.} \quad = \text{argmax}_m P(s_{L1}, s_{L2} | m)P(m) \\
 & \text{d.} \quad = \text{argmax}_m P(s_{L1} | m)P(s_{L2} | m)P(m)
 \end{aligned}$$

The scheme involves our conditional grammar models that assign a probability to surface realizations in a particular language, given an underlying meaning representation (or relative PMR), and an additional “semantic language model” for the prior probability of full PMRs. The combination step will require a heuristic search procedure similar to the decoder in machine translation. In section 4.3, we will discuss the advantages of applying this type of noisy channel model in parsing. The actual search in parsing is performed by a chart-based algorithm for parallel parsing (Kuhn 2005b).

⁹ The split of the translation problem in two subproblems leads to much more reliable results. Many of the strings receiving a high probability in a $P(e|f)$ model barely resemble English sentences, but since in the noisy channel model the probability is multiplied with the probability assigned by a language model (which in turn doesn't “know” anything about translation), the overall result is quite satisfactory.

¹⁰ The step from (4c) to (4d) makes the assumption that generating a sentence from a PMR in one language is independent from generating from the same PMR in other languages. This reflects the assumption that all relevant information is encoded in the PMR (which is of course a simplifying assumption, but an important one to make the approach practical).

3.3 Evaluation

Evaluating grammar induction systems (as opposed to treebank-trained grammars) is somewhat problematic; no acknowledged standard methodology exists (compare van Zaanen et al. (2004), who come to the same conclusion). Since in the initial PTOLEMAIOS architecture, the structural restrictions enforced by the PMR format convention are rather limited, this evaluation problem applies in part to our project as well. It is a research question for PTOLEMAIOS to identify adequate evaluation measures. For an appropriate set of measures, we will perform comparative evaluations for two variants of parsing: parallel parsing (with multilingual input) and simple parsing with monolingual input.

One set of evaluation measures commonly employed in work on grammar induction is the comparison of the induced grammar's behavior against an existing treebank. Besides such an evaluation against structural gold standard representations, we will work on a more task-oriented evaluation methodology. In collaboration with other projects, we will apply the induced grammars in multilingual information extraction or question answering tasks. Finally, a highly adequate application scenario, which at an advanced stage we will use for evaluation, is the use of parallel parsing of a corpus in order to improve statistical word alignment (and ultimately statistical machine translation). By exploiting the phrase correspondences of the most likely PMR, it can be expected that a simple word-based alignment can be improved.

4 Further discussion of the PTOLEMAIOS methodology

The PTOLEMAIOS project will develop a novel technique for training grammars on large amounts of corpus data. The technique requires only small amounts of hand-annotated text, hence it will be broadly applicable, even to languages for which no treebanks exist (and for which there is no high commercial interest in language technology, so one cannot expect costly treebanking activities in the future). Since the PTOLEMAIOS approach contains an indirect way of relating the observed surface forms in a language to the intended underlying form (by bootstrapping a consensus representation from a parallel corpus), it can be used for training linguistically sophisticated OT-style grammar models. The approach may thus open up new ways of exploring linguistic interactions of subtle, lexically grounded distinctions (e.g., ordering preferences, or the status of collocations and idiomatic expressions), for which large amounts of training data are required in order to make reliable predictions. In this section we provide some further background on some design decisions and argue for some of the expected advantages of the architecture.

4.1 Combination of insights from different fields

Our project combines the successful corpus-based methodology from NLP (including advanced machine learning techniques like the log-linear/Maximum Entropy models, which have turned out to be most adequate for linguistically more sophisticated models) with the formally well worked-out comparison-based architecture of OT, applying it to one of the most central topics in the cognitive sciences—language learning from limited input. Specifically, the OT insight is taken seriously that grammatical differences across languages are best modeled (and learned) by comparing possible ways of realizing the same underlying meaning representation: the language-specific conditional models for the realization of an underlying meaning m in language L_i (i.e., probability models for $P(s_{Li} | m)$) are the crucial grammar resources in our system.

To simulate the meaning-based optimization in learning experiments following “plain” OT, a corpus would have to be annotated with the appropriate meaning representations, and a generation grammar would have to be developed that produces a candidate set. Rather than pursuing this idea, the PTOLEMAIOS project uses the hypothesis that much of the annotation effort can be avoided if a (multilingual) parallel corpus is used as the empirical basis and a lean “pseudo meaning representation” (PMR) is induced by parsing the various languages in parallel. The grammars used in parsing are part of a bootstrapping cycle, i.e., the improved grammars learned from the parallel corpus will lead to a better PMR in the next cycle. Note that this way of applying bootstrapping comes close to the situation of the human learner who has to use the imperfect grammar at his/her present stage, along with language-independent information in order to learn the adult grammatical system.¹¹

The assumption of expressive optimization models as the language-specific grammar models will put us into a position to experiment directly with representations and sets of constraints assumed in the theoretical (OT) literature in linguistics.

4.2 Applicability of the PTOLEMAIOS grammars

Initially, the character of the underlying pseudo meaning representation is primarily determined by its function as the common part in a comparison of (cross-linguistic and language-internal) generation alternatives. However, its language-independent character will be an important factor for the broad usability of the PTOLEMAIOS grammars in various NLP tasks. The great advantage of the grammars resulting from induction with the PTOLEMAIOS system is that they produce parallel tree representations in parsing, i.e., they can be used directly in multilingual applications such as multilingual information

¹¹ *However, we make no claims that the PTOLEMAIOS architecture models the cognitive language acquisition process in humans in any direct way.*

extraction, question-answering, and machine translation (for example to improve statistical alignment, or in work on hybrid machine translation systems).

4.3 Exploitation of a noisy channel model in parsing

The use of a noisy channel model in parsing (i.e., using a combination of a “semantic language model” $P(m)$ and a conditional model $P(s|m)$ to determine the most likely parse/meaning representation m for a given string s) is largely unexplored.¹² Most statistical parsing approaches take a more direct approach, applying a model of $P(m|s)$ (or $P(m,s)$). When fully supervised learning is applied, the direct approach is of course very natural. However in a weakly supervised scenario building on an unannotated corpus, training has to live with meaning representations (m) that are not 100% reliable. In contrast, the actual sentence s is always known. Incorrectly labeled training data will thus always cause wrong predictions if we estimate and apply $P(m|s)$ directly, whereas we may be luckier if we estimate $P(s|m)$ and use it in a noisy channel model: $P(m)$ may be low for incorrect meaning representations, so the better alternative may indeed win. This presupposes of course that we have a broader empirical basis for estimating the semantic language model $P(m)$ (otherwise it would always suffer from the same labeling mistakes). The PTOLEMAIOS hypothesis is that the use of a multilingual parallel corpus will provide this additional empirical breadth. For training a model $P(m)$, there are various ways of extending the training data beyond what is usable for training the conditional model for a particular language: the consensus requirement may be relaxed, so sentences for which no high-confidence PMR can be obtained using, say 11 languages, but for which there is a consensus among 4 or 5 languages, could be used in training the language-independent model for PMRs. This may expand the empirical basis considerably, so we can use more fine-grained learning features, e.g., lexicalization features, which may be too sparsely instantiated for the language-specific grammar models.

4.4 A (more) realistic model of the interplay of cognitive systems

The envisaged noisy channel model approximates a very basic split of cognitive systems and information sources involved in human sentence processing: as an input sentence s in language L is parsed, seeking to obtain the correct meaning m for it that was intended by the speaker, (at least) two knowledge sources interact: (i) grammatical knowledge (**linguistic competence**) is applied (i.e., the language-specific grammar model $P(s|m)$), and (ii) the plausibility of the arising meaning representation m is checked, taking into account **contextual and encyclopaedic knowledge** and making inferences as required. In our architecture, part (ii) is modeled rather crudely by a context-independent statistical distribution $P(m)$, but note that the isolation of this group of knowledge

¹² But compare Miller et al. 1994.

sources is already a considerable conceptual advance over classical statistical parsers which indistinguishably accumulate grammar-specific knowledge, general world knowledge (reflected by certain patterns observed in the corpus) and domain-specific knowledge. Besides the advantage of being able to train the semantic language model on larger amounts of data, the conceptual split of models for different cognitive resources will facilitate error analysis and the improvement of the various subparts of the model.

The combination of knowledge sources (i) and (ii) leads to an infinite search space—any meaning whatsoever is a candidate for the correct m . This means that a heuristic search procedure is required: the decoder. This is the technological correspondence to (iii) the **performance system** in human sentence processing (which also doesn't cover the full search space of the competence grammar).

To close this section, it is important to point out (again) that we make no claims that the PTOLEMAIOS architecture models the cognitive language acquisition process in humans in any direct way. Nevertheless, to a certain extent the availability of corresponding versions of a sentence from the language learner's input in other languages may serve as a feasible simulation of the real learner's non-linguistic information about the utterance situation, using her/his other cognitive capabilities such as vision etc., which undoubtedly play a crucial role in the bootstrapping of grammatical knowledge.

5 Conclusion

In this paper, the system architecture and structural representation format that we plan to use in the PTOLEMAIOS project was outlined and discussed. The project will be of both theoretical and practical interest, as it addresses computational and representational issues in grammar learning and learnability, and it will lead to an implemented system for bootstrapping robust grammars for language-technological applications like (multilingual) information extraction, question answering and machine translation. For empirically oriented work in the OT framework, we hope to provide a methodology that allows one to train OT grammar models on larger amounts of text, which is indispensable for addressing non-trivial interactions among various phenomena.

References

- Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1):45–86.
- Bresnan, Joan, Shipra Dingare, and Christopher Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt and T. H. King (eds.), *Proceedings of the LFG 01 Conference*. CSLI Publications.

- Gale, William A., and Kenneth Ward Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Meeting of Annual the Association for Computational Linguistics*, pp. 177–184.
- Geman, Stuart, and Mark Johnson. 2002. Dynamic programming for parsing and estimation of stochastic unification-based grammars. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp. 279–286, Philadelphia.
- Gildea, Daniel. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03), Sapporo, Japan*, pp. 80–87.
- Grimshaw, Jane. 1991. Extended projection. Unpublished Manuscript, Brandeis University.
- Jäger, Gerhard. 2004. Maximum Entropy models and Stochastic Optimality Theory. Manuscript, University of Potsdam.
- Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In Jennifer Spenader, Anders Eriksson, and Östen Dahl (Eds.), *Proceedings of the Stockholm Workshop on 'Variation within Optimality Theory'. April 26-27, 2003 at Stockholm Univ. Sweden*, pp. 111–120.
- Koehn, Philipp. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Manuscript, University of Southern California.
- Kuhn, Jonas. 2001. Generation and parsing in Optimality Theoretic syntax – issues in the formalization of OT-LFG. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*, pp. 313–366. Stanford: CSLI Publications.
- Kuhn, Jonas. 2003. *Optimality-Theoretic Syntax—A Declarative Approach*. Stanford, CA: CSLI Publications.
- Kuhn, Jonas. 2004. Experiments in parallel-text based grammar induction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pp. 470–477.
- Kuhn, Jonas. 2005a. An Architecture for Parallel Corpus-based Grammar Learning. In Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schröder, and Petra Wagner (eds.), *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*, pp. 132–144, Frankfurt am Main. Peter Lang.
- Kuhn, Jonas. 2005b. Parsing Word-Aligned Parallel Corpora in a Grammar Induction Context. To appear in *Proceedings of ACL 2005 Workshop on Parallel Text*. Ann Arbor, Michigan.
- Miller, Scott, Robert Bobrow, Robert Ingria, and Richard Schwartz. 1994. Hidden understanding models of natural language. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994)*, pp. 25–32.
- Och, Franz Josef, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29: 19–51.

- Ratnaparkhi, Adwait. 1999. Learning to parse natural language with Maximum Entropy models. *Machine Learning* 34: 151–175.
- van Zaanen, Menno, Andrew Roberts, and Eric Atwell. 2004. A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation. In *Proceedings of the Workshop on the Amazing Utility of Parallel and Comparable Corpora, LREC 2004*, Lisbon, Portugal.
- Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23: 377–403.