Word Alignment Step by Step

Jörg Tiedemann

Department of Linguistics, Uppsala University

joerg@stp.ling.uu.se

Abstract

In this paper the current stage of the Uppsala Word Aligner (UWA) is described. The system is developed within the project on parallel texts, PLUG, which has its focus on the analysis of bi-lingual text collections with Swedish either as the source or the target language. UWA comprises a set of knowledge-lite approaches[1] for word alignment and lexicon extraction. A distinctive feature is its modularity. In the article, the main principles of the alignment software are introduced, different configurations and approaches are described, and examples of alignment results are presented.

1. Introduction

Word alignment aims at the identification of translation equivalents between linguistic units below the sentence level within parallel text (Merkel 1999), mainly bilingual text (*bitext*). Those units include single-word units (*SWUs*) and multi-word units (*MWUs*) and will be referred to as *link units* further on. The basic terminology for describing parallel text and word alignment in this paper follows Ahrenberg et al (1999) and Ahrenberg et al (forthcoming). In particular, each word correspondence in the bitext describes a *link instance*, or simply a *link*. A pair of link units that is instantiated in the bitext will be referred to as *link type*. Word alignment systems usually assume segmented bitext (*sentence aligned bitext*). Common *bitext segments* are sentence fragments, sentences, and sequences of sentences that have corresponding units in the translation.

Depending on its purpose, a word alignment system attempts to maximize the number of discovered links (→ word instance alignment) (e.g. Ahrenberg et al 1998, Melamed 1999) or the number of extracted link types (→ bilingual lexicon extraction) (e.g Melamed 1995, Resnik and Melamed 1997, Tiedemann 1998a). Lexicon extraction aims at providing correct translations whereas word alignment has to deal with insertions, deletions, and other modifications within the bitext as well. Furthermore, word alignment systems may focus on specific types of link units, e.g. terms (Dagan and Church 1994, van der Eijk 1993) and collocations (Smadja et al 1996).

The task of word alignment is not trivial especially because it goes beyond simple one-to-one word correspondences in many cases. Multi-word units (MWUs) have to be handled due to the use of non-compositional compounds, associated idiomatic expressions, multi-word names and so on. The difference in compounding between different languages increases the difficulties with the identification of appropriate

Proceedings of NODALIDA 1999, pages 216-227

correspondences further. In addition, the text type is decisive for the word alignment process. Technical text tends to include specific terms and simple structures that are translated directly while e.g. literary texts include many language-specific idioms.

Concurrently, Martin Kay's proposal for approaching machine translation can be applied to word alignment as well:

"The keynote will be modesty. At each stage, we will do only what we know we can do reliably. Little steps for little feet!" (M.Kay 1980)

The alignment of MWUs can be approached in different ways. Smadja et al (1996) propose the compilation of source language collocations using statistical co-occurrence measures (*static segmentation*). The appropriate correspondent is found by iterative extension of the link unit in the target language segment (*dynamic segmentation*). Another approach applies collocation lists for both languages, which have been compiled in advance from the bitext (Ahrenberg et al 1998, Tiedemann 1998). MWUs are then handled like single tokens for both languages. A third possibility is to expand link units iteratively for both languages in order to find the most appropriate link. Melamed (1997) uses iterative processing in order to optimize the underlying translation model. The iteration is alternated in order to cover MWUs in both languages.

The word alignment system, which is introduced in this paper, supports all the three approaches to text segmentation as far as contiguous phrases are concerned. The approach to dynamic segmentation differs from Melamed in the usage of ranked candidate lists instead of translation models. Furthermore, classified stop word lists are used for improving the result and reducing the search space.

2. The Uppsala Word Aligner (UWA)

The Uppsala Word Aligner is developed within the co-operative project on parallel texts, PLUG[2] (Sågvall Hein, forthcoming). The goal of PLUG is to develop, apply, and evaluate software for the alignment and generation of translation data. Word alignment is one of the major issues at hand. UWA is based on earlier studies on bilingual lexicon extraction (Tiedemann 1997, 1998a). It combines several knowledge-lite approaches to word alignment. The system is integrated into the Uplug toolbox (Tiedemann forthcoming), which provides a convenient environment for the work with modular corpus tools.

As mentioned earlier, word alignment walks with small feet. Therefore, the proposal is to combine different approaches, to collect available knowledge sources, and to reach the goal by little steps.

The principles of baby-steps

1. Prepare carefully before taking the first step.
2. Use all available tools that can help.
3. Check alternatives before taking the next step.
4. Take safe steps first; try risky steps later.
5. Remove everything that is in the way.
6. Improve walking by learning from previous steps.
7. Reach the goal with many steps rather than one big one.
8. Continue trying until you cannot get closer.

Based on these general principles, UWA was designed as a modular and iterative (rule 6+8) system. The bitext runs through initial pre-processing steps before the alignment starts (rule 1). Alignment candidates are collected from any appropriate source (rule 2). Candidates are ranked by their reliability, e.g. association scores (rule 3). The most reliable candidate is aligned first (rule 4). The alignment process is split into a sequence of separated steps (rule 7). Aligned link units are removed from the search space (5).

In the following the three main phases of the UWA are described: text segmentation, candidate collection, and alignment of link units.

2.1 Text Segmentation

UWA assumes sentence-aligned bitexts. However, an initial sentence alignment step can be added.

UWA provides modules for the work with static and dynamic text segmentation. In the case of dynamic segmentation, the text will be simply tokenized, i.e. segmented into SWUs and punctuation marks. In case of static segmentation, this phase accounts for a subsequence segmentation of the bitext into link units. It includes tokenization, the recognition of MWUs, and the actual segmentation of the text into link units. The recognition of MWUs can be automated. UWA applies iterative processing for the compilation of word collocations. The association between word units and their subsequent words is measured in terms of mutual information scores. As proposed in Ahrenberg et al (1998), classified lists of functional words are used to reduce search space and to exclude ungrammatical constructions. Consider the small example of classified stop words for English phrase generation, which is illustrated in figure 1.

Figure 1: Classified stop word lists (lower case).

```
skip token = '((,[,{,),],},\,',!,?)'
skip at = '(or,and,but,not)'
skip before = '(i,you,he,she,it,we,they)'
skip after = '(mine,yours,his,hers,its,ours,theirs)'
non-phrase-starter = '(my,your,his,her,our,their)'
non-phrase-ender = '(the,a,an)'
skip at string type = '(numeric)'
```

Stop words are divided into 6 types. '*Skip token*' items are not considered at all in any segmentation. Furthermore, the segmentation will stop at '*skip at*' tokens. They are considered to be single word units and the segmentation process continues with the subsequent token. '*Skip before*' defines link unit breaks in front of each instance of each word that is specified. Similarly, '*skip after*' defines breaks after each instance of words in the list. '*Non-phrase-starter*' and '*non-phrase-ender*' are not allowed in the beginning or at the end of any phrase, respectively. However, those words may appear within phrasal constructions as e.g. in 'in my mind' or 'in a row'. Note, that the definite article is allowed in the beginning of a phrase. In this way correspondences between definite forms of English and Swedish nouns can be recognized[3]. Furthermore, each category may include all tokens of a certain string type. In the example above, all numeric tokens will be added to the 'skip at' list. In cases of overlapping definitions the stronger restriction is chosen. In the current stage of the system only four of the classes above are used: 'skip token', 'skip at', 'non-starter', and 'non-ender'.

In figure 2 a sample of an automatically generated list of English collocations is presented. It is based on an English subtext from the PLUG corpus (Tiedemann 1998b) with about 66,000 words. The minimal frequency was set to 4.

Figure 2: Generated phrases with frequency>4 (case-folded).

MI	freq	collocation
10.039	4	the yom kippur
10.039	4	raymond aron
10.039	4	danny kaye
9.717	5	yom kippur
9.717	4	the golan heights
9.717	4	golan heights
9.454	6	world war ii
9.454	6	the mishkenot sha
9.454	4	lyndon johnson
9.454	4	american public opinion
9.395	4	justice cohn
9.231	7	tel aviv
9.231	7	mishkenot sha
9.231	5	the ottoman empire
9.231	5	ottoman empire

In the current stage, the system provides contiguous phrases only. Static text segmentation applies a simple left-to-right process. It starts with the left-most token in the bitext and looks for the longest valid link unit. The segmentation continues to the right of the last validated link unit until the complete bitext is processed. Here, single word units always represent valid link units and sentence breaks always mark the end of the current link unit.

2.2 Identification and Collection of Candidate Pairs

In this part the system compiles and collects translation equivalents. Sources are pre-compiled collections and generated lists of candidate pairs. In the current implementation, UWA applies the following sources:

- pairs of associated word units (applying co-occurrence measures)
- cognate lists (applying string similarity measures)
- single word bitext segments
- pairs of low frequency units
- machine readable bilingual dictionaries (MRBDs)
- previously aligned word pairs (iteration)

Collections of candidate pairs are compiled by investigations on the association between link units. UWA applies co-occurrence measures and string similarity scores in order to find alignment candidates. The number of possible candidates is reduced by some general restrictions in order to improve the performance:

link distance: link units have to occur within a certain distance between their positions in the bitext segment
string length: each link unit has to exceed a minimal length
length difference ratio (LDR): the ratio between the length of the shorter link unit and the length of the longer link unit has to pass a certain threshold value
string type: each link unit has to present a certain string type (e.g. 'contains at least one alphabetic character')
frequency: the number of occurrences of each link unit has to exceed a certain value (for co-occurrence measures only)
co-occurrence frequency: each pair of link units has to co-occur at least a certain number of times (for co-occurrence measures only)

The value of each of the parameters above can be adjusted to the type of investigation in progress. Certainly, string length and LDR should be restricted for investigations on string similarity, whereas frequency thresholds are important for co-occurrence measures.

UWA supports three *word association* scores: the Dice coefficient, mutual information, and t-score. The current investigations were focused on the application of the Dice coefficient.

$$Dice = \frac{2\,prob(S,T)}{prob(S) + prob(T)}$$

In our case S and T represent the link units in the source and the target language under consideration. The probabilities of S and T to occur in the text, and the probability of both units to co-occur in the same bitext segment (sentence alignment) can be estimated by appropriate frequency counts. Simple stemming functions are used in order to reduce the inflectional variety of words in different languages and to improve the statistical calculations.

String similarity can be measured by different metrics (Melamed 1995, Borin 1998). UWA uses the Longest Common Subsequence Ratio (LCSR). UWA applies dynamic programming for computing the length of the longest common subsequence (LCS) of two strings (Stephen 1992). This value, divided by the length of the longer string, provides a measure for string similarity between them. In figure 2, the LCSR calculation is illustrated. In the figure, the application of the algorithm with MWUs is demonstrated as well.

Figure 2: The longest common subsequence ratio of 'see example' and 'se exempel'.

	s	e	e		e	x	a	m	p	l	e
s	1	1	1	1	1	1	1	1	1	1	1
e	1	2	2	2	2	2	2	2	2	2	2
	1	2	2	3	3	3	3	3	3	3	3
e	1	2	3	3	4	4	4	4	4	4	4
x	1	2	3	3	4	5	5	5	5	5	5
e	1	2	3	3	4	5	5	5	5	5	5
m	1	2	3	3	4	5	5	6	6	6	6
p	1	2	3	3	4	5	5	6	7	7	7
e	1	2	3	3	4	5	5	6	7	7	8
l	1	2	3	3	4	5	5	6	7	8	8

$$LCSR = \frac{length[LCS(S_1, S_2)]}{\max[length(S_1), length(S_2)]} = 8/11 \approx 0.72$$

Further investigations on string similarity metrics have been carried out (Tiedemann 1999) but they have not yet been applied in the word alignment process.

Another source of alignment candidates can be found in *single word bitext segments.* UWA considers each bitext segment with exactly one link unit in one language to be a valid alignment candidate.

Low frequency link units cannot be recognised by statistical association scores. However, they represent a large portion of general text corpora. UWA applies a simple heuristic in order to extract alignment candidates of low frequent text units. Assuming two frequency thresholds t_1 and t_2 with $t_2 < t_1$, the system removes all units that occur less than t_1 times in the complete text from each bitext segment. Now, each bitext segment with exactly one remaining link unit on each side is considered to be a valid alignment candidate, if both link units occur less than t_2 times. However, finding appropriate values for t_1 and t_2 is not trivial. Mainly, the distance between t_1 and t_2 is significant for the quality of extracted pairs.

Furthermore, *MRBDs* of any origin can be added to the collection of alignment candidates. Certainly, their quality is decisive for the quality of the word alignment. This includes that the chosen MRBDs should be suitable to the type of text under consideration.

As mentioned earlier, UWA supports dynamic text segmentation. If this alternative is chosen, the system generates all combinations of possible link units. The number of possible units can be reduced drastically by the use of classified stop word lists. The same principles as for static text segmentation are applied (except co-occurrence thresholds). Consequently, only contiguous phrases are identified. Association scores are computed for each possible link unit combination. This includes co-occurrence measures as well as string similarity measures. In this way, a list of alignment candidates is collected that includes all link unit combinations that pass a certain threshold value.

2.3 Word Alignment

The actual word alignment is based on the previously collected alignment candidates. Each bitext segment runs through a sequence of alignment steps. Candidates of word instance alignments can be compiled by associating possible link units. These units may include parts of the static segmentation or may be compiled dynamically. The same restrictions as in the candidate collection phase are applied in order to reduce the search space.

The alignment starts with the most reliable candidates. Each aligned token is removed from the text and only non-aligned tokens remain for the next step. In the current version of the system, 9 alignment steps are defined:

1. align one token units
2. align identical numerics
3. align cognates (string similarity scores, high threshold)
4. align strongly associated units (co-occurrence measures)
5. align low frequency pairs
6. align pairs from MRBDs
7. align cognates (string similarity scores)
8. align associated units (co-occurrence measures)
9. align remaining one token units

Each alignment step can be adjusted by several parameters such as LDR, link distances, string length and type. The alignment candidates are ranked by their probability (if an appropriate value is defined, e.g. Dice scores) and the most reliable pairs will be aligned first. Position weights can be used to modify probabilistic scores. The distance between the actual position and the estimated position of the aligned link unit, multiplied with a certain factor, is used to reduce the association score of each candidate pair. The reduction factor can be adjusted for each alignment step separately.

2.4 Iteration

Previously aligned word pairs that have been removed from the text can be added to the collection of alignment candidates (principle 6). The iteration process can be described as follows:

1. compile a bilingual lexicon from previously aligned words
2. compute alignment candidates by means of word association scores using the remaining tokens in the corpus
3. start the word alignment process all over again including an additional alignment step that applies the newly compiled lexicon
4. continue with (1) until no new alignments can be found (principle 8)

2.5 Evaluation

UWA stores information about each link. Each aligned unit is represented by a unique identifier corresponding to its origin in the bitext and its byte-span within the text relative to the beginning of the bitext segment. Reference data in form of a "gold standard" were manually defined for each bitext under consideration. UWA includes an evaluation module that compares results from a word alignment process with the gold standard. The module produces a protocol with information about each pair from the gold standard and summarizes the alignment result by counting the number of correct, partially correct, incorrect, and not aligned pairs. Finally, evaluation metrics are calculated. In this paper, $recall_{PWA}$, $precision_{PWA}$, and F-measure as their geometric mean were applied as proposed in Ahrenberg et al (forthcoming). Furthermore, information about the actual alignment step is stored for each aligned pair. In this way, the alignment process can be retraced and the quality of each step be investigated.

3. Experiments

UWA was tested with English/Swedish and Swedish/German bitexts from the PLUG corpus. For an illustration, word alignment results from an English→Swedish sub-corpus are presented here. The bitext is a literary text of some 130,000 words[4]. The gold standard comprises 500 random source language units that were linked manually using the Plug Link Annotator (Merkel et al forthcoming). In table 1, results of several word alignment experiments with different UWA settings are summarized.

	$precision_{PWA}$ (%)	$recall_{PWA}$ (%)	F (%)	time (min)[5]
static, no iteration	82.28	32.00	46.09	70
static	78.73	42.22	54.96	108
static+steps	79.88	43.91	56.67	115
semi-dynamic	77.34	43.05	55.32	131
semi-dynamic +steps	78.84	44.77	57.11	131
dynamic	78.73	43.80	56.28	250
dynamic +steps	78.27	44.29	56.57	261
parameter optimisation +MRBD	83.51	51.61	63.80	132

Table 1: Word alignment results for different UWA configurations.

The differences between each alignment configuration need to be explained. The first experiment represents the most basic configuration without iteration. The first three experiments apply *static* text segmentation only. Dynamic segmentation was applied in calculating string similarity measures and in the word alignment phase for the alignment

experiments that are denoted with *semi-dynamic*. The two *dynamic* approaches apply additionally dynamic text segmentation for the compilation of associated link candidates. The term *+steps* indicates the use of all alignment steps as described in section 2.3. The other experiments, except the last one, apply a simplified alignment procedure with one step for each candidate collection only. In the last approach, an empirically optimized UWA configuration and an additional basic dictionary were applied. Here, semi-dynamic text segmentation was used.

The results in table 1 confirm the practical use of the alignment principles that were described above. The approaches that apply a fine-tuned sequence of alignment steps (+step) contribute to the performance (in terms of F-values) at almost no expense (considering e.g. the processing time). Furthermore, the combination of static and dynamic text segmentation seems to be the most worthwhile approach. However, the set of alignment parameters has to be investigated further in order to discover potential improvements.

4. Conclusions

The Uppsala Word Aligner represents a highly adjustable word alignment system for fast and robust alignment of words and contiguous phrases from bilingual parallel texts. It supports different configurations and parameter settings for systematic investigations on translation units below the sentence level. The system applies knowledge-lite approaches that can be adjusted to different language pairs easily. Furthermore, supplementary modules and knowledge sources can be added. UWA is integrated in a modular corpus toolbox that provides convenient tools for experimentation, generation, and data organisation. It also includes a module for automatic evaluation using a previously defined gold standard. Thus, empirical investigations of different approaches and configurations are supported in a very efficient way.

UWA is implemented mainly in Perl and was tested on Linux. It will be available for academic research purposes at the end of the PLUG project.

[1] Knowledge-lite approaches in this case comprise techniques with minimal linguistic resources needed.

[2] The PLUG project is jointly funded by "The Swedish Council for Research in the Humanities and Social Sciences" HSFR and "The Swedish National Board for Industrial and Technical Development" NUTEK.

[3] The definite form of Swedish nouns is created by morphological modification.

[4] The word count includes both languages.

[5] The processing time includes phrase generation (8:42 min) and cognate extraction (23:54 min for static segmentation and 39:59 min for dynamic segmentation). However, these data are reusable and therefore, do not need to be compiled again for each alignment experiment.

References

Ahrenberg, L., Andersson, M. and Merkel, M. 1998. A simple hybrid aligner for generating lexical correspondences from parallel texts. In *Proceedings of COLING-ACL'98*, Montreal, Canada, pp. 29-35.

Ahrenberg, L., Merkel, M., Sågvall Hein, A., and Tiedemann, J. 1999. Evaluating LWA and UWA. PLUG deliverable 3A.1. Internal report.

Ahrenberg, L., Merkel, M., Sågvall Hein, A., and Tiedemann, J. forthcoming. Evaluation of Word Alignment Systems. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC-2000*, Athens, Greece, 2000.

Borin, L. 1998. Linguistics isn't always the answer: Word comparison in computational linguistics. In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALI98*, Center for Sprogteknologi and Department of General and Applied Linguistics, University of Copenhagen, pp. 140-151.

Dagan, I. and Church, K. W. 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, Stuttgart/Germany.

van der Eijk, P. 1993. Automating the Acquisition of Bilingual Terminology. In *Proceedings of the 6th Conference of the European Chapter of the ACL*, 1993, Utrecht/The Netherlands.

Kay, M. 1980. The Proper Place of Men and Machines in Language Translation. Xerox PARC Working Paper, reprinted in *Machine Translation 12 (1-2)*, 1997, pp. 3-23

Melamed, I. D. 1995 Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons. In *Proceedings of the 3rd Workshop on Very Large Corpora*, Boston/Massachusetts.

Melamed, I. D. 1997. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, Providence.

Melamed, I. D. 1999. Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics, 25(1)*, pp. 107-130.

Merkel, M., Andersson, M., and Ahrenberg, L. forthcoming. The PLUG Link Annotator - Interactive Construction of Data from Parallel Corpora. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University.

Merkel, M. 1999. *Understanding and enhancing translation by parallel text processing.* Linköping Studies in Science and Technology. Dissertation No. 607. Linköping University. Dept. of Computer and Information Science.

Resnik, P. and Melamed, I. D. 1997. Semi-automatic acquisition of domain-specific translation lexicons. In *Proceedings of the Conference on Applied Natural Language Processing*, Washington, D.C.

Smadja, F., McKeown, K. R., Hatzivassiloglou, V. 1996. Translation Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics, 22(1).*

Stephen, G. A. 1992. String Search. Technical report TR-92-gas-01, School of Electronic Engineering Science, University College of North Wales, Gwynnedd.

Sågvall Hein, A. forthcoming. The PLUG Project: Parallel corpora in Linköping, Uppsala, Göteborg: Aims and achievements. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University.

Tiedemann, J. 1997. Automatical Lexicon Extraction from Aligned Bilingual Corpora. Diploma thesis, Otto-von-Guericke-University, Magdeburg, Department of Computer Science.

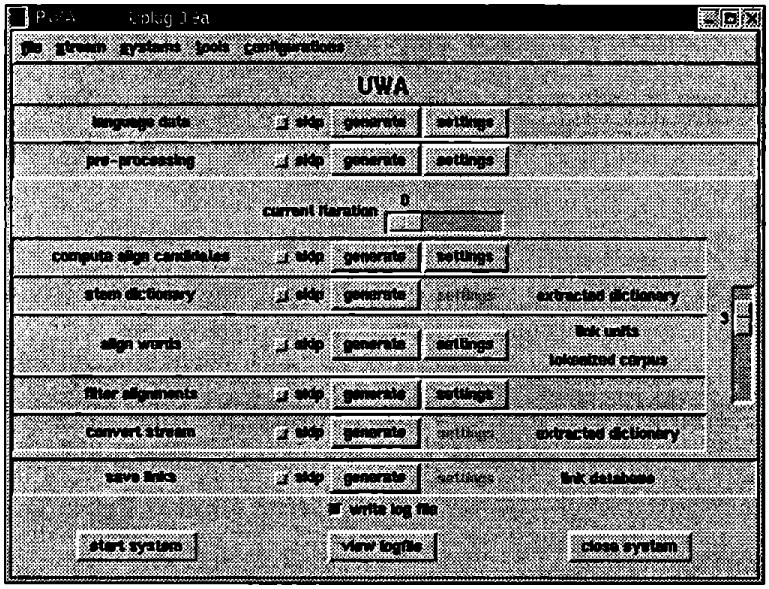
Tiedemann, J. 1998a, Extraction of translation equivalents from parallel corpora, In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALI98*, Center for Sprogteknologi and Department of General and Applied Linguistics, University of Copenhagen, pp. 120–128.

Tiedemann, J. 1998b. Parallell corpora in Linköping, Uppsala and Göteborg (PLUG). Work package 1. PLUG report, Department of linguistics, Uppsala university.

Tiedemann, J. 1999. Automatic Construction of Weighted String Similarity Measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Coropora*, University of Maryland, College Park/MD, pp. 213-219.

Tiedemann, J. forthcoming. Uplug - A Modular Corpus Tool for Parallel Corpora. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University, Sweden.

Appendix A: UWA in the Uplug environment



Appendix B: MWU links from the English/Swedish test corpus (random sample)

source	target
Kippur War	Kippur-kriget
The Prime Minister	Premiärministerns
Kippur attack	Kippur-anfallet
Kosher food	Koschermat
Kultur paradise	Kultur-paradis
capitalist democracy	kapitalistiska demokratierna
careers be	karriärer
central tragedy	centrala tragedi
did mean	betydde
The Soviet Union	Sovjetunionen
anything	vad som helst
centralized state capitalism	centraliserad statskapitalism
The Times	New York Times
chamber music	kammarmusik
The United States	The United States
choke points	chokepunkterna
The West Bank	Västbanken
circus style	cirkusstil
The White House	Vita huset
citrus green	citrusgrönska
The Zionist movement	Sioniströrelsen
citrus groves	citruslundarna