ID	Lemma	POS	Start	End	Sentence
test.en-ru.18	light	NOUN	46	51	Using a technique for concentrating the solar **light**, resulted in an overall efficiency of 20%.
			39	50	Каждый представитель может выступать в **зависимости** от полученных указаний.
test.en-ru.19	light	NOUN	46	51	Using a technique for concentrating the solar **light**, resulted in an overall efficiency of 20%.
			2	8	С **учетом** работы, оратор считает целесообразным изложить принципы.

Table 4: Excerpt from the cross-lingual dataset (En-Ru): two sentence pairs sharing the same first sentence are shown, with the target word occurrence in bold type.

all sentences are manually selected and annotated, and that Arabic and Russian are included in a Word-in-Context dataset for the first time.

We report two cross-lingual instances (sentence pairs) in Table 4 for the En-Ru language combination, which share the first sentence. Given the English lemma *light*, its part of speech (noun), and two sentences, one in English where *light* occurs and one in Russian where a translation of *light* appears, participants are asked to determine whether the target occurrence (in bold in the Table) of *light* and its translations into Russian зависимости and учетом share the same meaning or not. Importantly, translations are allowed to be multi-word expressions and periphrases.

The cross-lingual sub-task comprises test data only and includes 500 unique English lexemes and 1000 sentence pairs for each language combination as reported in Table 1 (bottom). Note that, in this case, all cross-lingual datasets share the same English target lexemes. Similarly to its multilingual counterpart, the data in this sub-task contains a balanced number of T (50%) and F (50%) tags.

3.3 Selection of the data and annotation

Sources of the data In order to construct MCL-WiC, we leveraged three resources. First, we used the BabelNet[4] multilingual semantic network (Navigli and Ponzetto, 2010) to obtain a set of lexemes in all languages of interest. Subsequently, we extracted sentence pairs containing occurrences of such lexemes from two corpora, namely the United Nations Parallel Corpus (Ziemski et al., 2016, UNPC)[5] and Wikipedia[6]. UNPC is a collection of official records and parliamentary documents of the United Nations available in the six UN languages[7], whereas Wikipedia is a wide-coverage multilingual collaborative encyclopedia. These corpora were selected due to their wide coverage in terms of domains and languages. In fact, such heterogeneity allowed for the creation of a new competitive benchmark capable of evaluating the generalization ability of a system in discriminating senses in different domains and across languages. With this aim in view, we derived 50% of the selected sentence pairs from UNPC and the remaining 50% from Wikipedia.

Selection of lexemes Starting from BabelNet, we extracted a set of 5250 unique ambiguous lexemes in English and 1000 unique lexemes for each of the following languages: Arabic, Chinese, French and Russian. The selected pairs in English were distributed as follows: 4000 for the training data, 500 for the development data and 750 for the test data (500 for the multilingual sub-task and 250 for the cross-lingual sub-task[8]; we enriched the latter with additional 250 pairs derived from the multilingual test data). Instead, the selected pairs in languages other than English were included in the multilingual sub-task only and distributed as follows: 500 for the development data and 500 for the test data. We selected the target lexemes starting from basic vocabulary words and such that they had at least three senses in BabelNet. A key goal was to cover all open-class parts of speech, namely nouns, verbs, adjectives and adverbs, whose distribution in MCL-WiC is shown in Table 5. The target lexemes were chosen so as to avoid phrasal verbs and multi-word expressions.

[4] https://babelnet.org/
[5] https://conferences.unite.un.org/uncorpus/
[6] https://wikipedia.org

[7] Arabic, Chinese, English, French, Spanish and Russian.
[8] We recall that, in the cross-lingual sub-task, the target lexemes are provided in English and shared across all datasets.

	En-En			Ar-Ar		Fr-Fr		Ru-Ru		Zh-Zh		En-*
	Train	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Test
NOUN	4124	582	528	490	494	548	514	572	582	520	554	458
VERB	2270	246	298	428	398	262	272	352	372	330	364	320
ADJ	1430	158	144	72	98	156	184	54	30	122	62	178
ADV	176	14	30	10	10	34	30	22	16	28	20	44

Table 5: Part-of-speech distribution in MCL-WiC. * indicates all languages supported in MCL-WiC other than English.

Selection and annotation of sentence pairs For each of the target lexemes, we annotated two sentence pairs from either UNPC or Wikipedia. All selected sentences were well-formatted and, most importantly, provided a sufficient semantic context to determine the meaning of the target occurrences unequivocally. Subsequently, each sentence pair was associated with a tag, depending on whether the target words in the two contexts are used with the same meaning (T) or not (F). To perform both the selection of the data as well as the annotation, we employed eight annotators with a high level of education and linguistic proficiency in the corresponding language; the annotation work required approximately six months. Importantly, all annotators followed specific criteria which we describe in the following paragraph.

Annotation criteria We provided each annotator with general annotation guidelines. Besides general criteria, each annotation team[9] established ad-hoc guidelines for specific linguistic issues, some of which will be briefly illustrated in Section 4, below.

General annotation criteria can be broadly divided into grammatical and lexicographic-semantic criteria. The former refer to the format and the grammatical correctness of the sentences to be selected: annotators were asked to choose well-written sentences only, i.e. sentences with a clear structure, ending with a full stop and containing a main clause. Instead, lexicographic-semantic criteria refer to the attribution of the labels. To determine whether two occurrences were used with the same meaning or not, annotators were asked to use multiple reputable dictionaries (e.g. for English we used the Merriam-Webster, Oxford Dictionary of English and English Collins dictionaries). Moreover, to avoid misperceptions in the same-sense tagging annotations, we asked annotators to justify their choices by providing substitutes for the target occurrences with synonyms, hypernyms, paraphrases or the like. Contrary to what was done in WiC and XL-WiC, we argue that, for the purposes of this task, annotating according to lexicographic motivations, i.e. by using reliable dictionaries, contributes significantly to minimizing the impact of subjectivity, thus producing more adequate and consistent data. Finally, lexicographic-semantic criteria also provided concrete indications and examples regarding the attribution of tags. For instance, T was used if and only if the two target occurrences were used with exactly the same meaning or, in other words, if, using a dictionary, the definition of the two target words was the same.

Inter-annotator agreement In order to determine the degree of uncertainty encountered during the annotation process, we computed the inter-annotator agreement. To this end, we randomly selected a sample of 500 sentence pairs from each of the En-En and Ru-Ru multilingual datasets, and 200 sentence pairs from the En-Ar and En-Zh cross-lingual datasets. Validators were provided with the same guidelines used during the annotation process. We calculated the agreement between two different annotators using the Cohen's kappa, obtaining κ=0.968 in En-En, 0.952 in Ru-Ru, 0.94 in En-Ar and 0.91 in En-Zh, which is interpreted as almost perfect agreement.

Data format For each sub-task, we provide two types of file (.data and .gold) in JSON format. The .data files contain the following information: a unique ID, the lemma, its part of speech, the two sentences and the positional indices to identify the target occurrences to be considered (see Tables 2 and 4). Instead, the .gold files include the gold answers, i.e. the corresponding ID and tag, as shown in Table 3.

[9] An annotation team is made up of annotators working on the same language.

4 Linguistic Issues

In this section, we describe interesting language-specific issues which required additional guidelines. Due to space limits, we focus on languages which do not use the Latin alphabet, i.e. Arabic, Chinese and Russian, illustrating only the most significant issues encountered.

Arabic From a WSD perspective, compared to other languages, written Arabic poses bigger challenges due to the omission of vocalization, which increases the degree of semantic ambiguity. In fact, the vocalization, expressed by diacritics placed above or below consonants, contributes significantly to determining the right interpretation and thus the meaning of words. For instance, the unvocalized word form *b-r-d* could be interpreted as *bard* ("cold"), *burd* ("garment") or *barad* ("hail"). Of course, in Arabic, polysemy also affects vocalized words, which can have multiple meanings, e.g. *ummiyy* means "maternal", but also "illiterate". For the purposes of MCL-WiC, we chose to keep the sentences as they are found in UNPC and Wikipedia, i.e. unvocalized in the vast majority of cases, while – instead – providing the target lemmas in the vocalized form. This was done in order to avoid lexical ambiguity deriving from lemmas which share the same word form but are vocalized in a different way. Furthermore, this choice facilitated the selection and annotation of sentence pairs in which a given target lemma occurs.

Chinese Since Chinese does not adopt an alphabet, the semantic ambiguity that can be found in English homographs is basically lost. In Chinese, if two unrelated words are pronounced in the same way, such as "plane" (the airplane) and "plane" (the surface), they are not usually written in the same way. By way of illustration, 沉默, meaning "silent; to be silent" and 沉没, "to sink", are both pronounced as *chénmò*, but, because they are written with different characters, they cannot be considered ambiguous words. Analogously, some characters have an extremely high semantic ambiguity themselves, but since they appear most frequently in polysyllabic words, their ambiguity is lost. For example, the character *guǒ* 果 has at least two meanings, "fruit" and "result", but this character almost never stands as a word on its own in contemporary Chinese. In the current lexicon most of the Chinese words are composed of two or more characters; when it appears in actual texts, *guǒ* is almost always connected to other characters, and the word thus formed is no longer semantically ambiguous. Finally, similarly to the cross-lingual sub-task, some ambiguity had to be discarded in translation, as in the case of Chinese classifiers which have a marked potential for semantic ambiguity. For example, *dào* 道 is, among others, the classifier for long and narrow objects, as in *yī dào hé* 一道河, a river (one+classifier+river), or for doors, walls and similar objects with an entry and an exit, as in *yī dào mén* 一道门, a door (one+classifier+door). However, since classifiers are virtually absent in European languages, they could not be applied in the cross-lingual sub-task and were discarded.

Russian A noteworthy issue encountered by Russian annotators concerned the verbal aspects which can be viewed as one of the most challenging features of the Russian language especially for L2-learners[10] with no Slavic background. In Russian, a verb can be perfective, imperfective or both. Normally, a perfective verb has one or more imperfective counterparts and vice versa. Broadly speaking, perfective verbs are typically used to express non-repetitive actions completed in the past, or actions which will certainly be carried out in the future, and also in general for past or future actions for which the speaker intends to emphasize the result that was or will be achieved. Conversely, imperfective verbs are used to express actions which are incomplete, habitual, in progress, or actions for which the speaker does not stress the result to be attained. In MCL-WiC, given a verbal target lexeme, we decided to choose sentences in which the target words occurring in the selected sentences and the target lemma shared the same aspect. In fact, in Russian, although pairs of perfective and imperfective verbs such as делать, сделать (to do) or спрашивать, спосить (to ask) show a high degree of morphological relatedness, they tend to be considered as distinct lemmas.

Another interesting issue regards participles. In some cases, annotators raised issues concerning the part of speech of participles occurring as target words in the selected sentences. In fact, Russian participles derive from verbs, but are declined and can behave as adjectives. Since the target lexemes and the corresponding occurrences must share the same part of speech, we decided to discard sentences in which the part of speech of the target

[10] In language teaching, L2 indicates a language which is not the native language of the speaker.